FP7-212214

Title	**Data Preservation in the Social Sciences: Recommendations for a CESSDA Research Infrastructure (D10.4)**
Work Package	WP10
Authors	Heiko Tjalsma (ed.), Valentijn Gilissen, Henk Koning and Dirk Roorda (DANS)
Source	CESSDA-PPP Survey 2008
Dissemination Level	PU (Public)

Summary/abstract

This is the D10.4 “Report on data formats and proposals for file format registry.”

Contents

Executive summary

This report is an answer to the questions posed in work package 10, tasks 10.3 and 10.7, of the CESSDA-PPP: "Examine long-term preservation issues posed by multiple data formats and make recommendations relating to the construction and maintenance of a CESSDA-wide file format registry" and "Examine the potential impact on social science research of the expanding number of data file types and information formats, and critically assess the measures that an upgraded CESSDA RI (Research Infrastructure) will need to take in order to adequately acquire, preserve and disseminate these new forms of data".

As these two questions are strongly related they are answered in one, combined, report.

Chapter 2 focuses on the expected changes in the use of file formats and tools in the social sciences, mainly based on the survey carried out in 2008 among the CESSDA data archives. Two different changes are foreseen: one is the growing diversification of data formats, in particular qualitative data, and another is the diminishing position of statistical software programs as a standard for using and *storing* data. An important observation, based on the survey, is that the present state regarding digital preservation within CESSDA can be described as worrying. The implications for the CESSDA Research Infrastructure (RI) are being analysed by looking into the usefulness of general (reference) models like the OAIS and trusted digital repositories. Two conditions should be fulfilled to enable the partners within the CESSDA RI to function as trusted digital repositories:

1. The construction of a clear set of guidelines, "CESSDA-ERIC requirements for operating trusted data repositories": implying a CESSDA-ERIC 'seal of approval'.
2. Assessment procedures for CESSDA RI partners, appropriate to the level of service which the partner is going to provide.

In chapter 3 long-term preservation is discussed in a broader context and its main issues looked into. A critical assessment of existing preferred file formats (global format registries) and tools is given. The question of what preferred file formats are and how useful they are is answered.

In chapter 4 recommendations are formulated on the construction and maintenance of a CESSDA-wide file format registry and a preferred formats list. In particular attention is paid to the long term preservation of SPSS, SAS or STATA data files. The relation of a CESSDA-wide file format registry with other global registries is looked into. Attention is paid to tools for recognition and conversion of file formats.

The appendices contain lists of preferred format lists, conversion tools and file format identification tools.

A list summarising the recommendations of this report can be found at the end.

1. Introduction

The aim of this report is to answer the questions posed in work package 10, tasks 10.3 and 10.7, of the CESSDA-PPP: the Preparatory Phase Project for a Major Upgrade of the CESSDA Research Infrastructure (RI). These questions focus on the long term digital preservation of the research data, as well as their documentation, in the care of the CESSDA organisations. More specifically, the assigned tasks in the project plan were "Examine long-term preservation issues posed by multiple data formats and make recommendations relating to the construction and maintenance of a CESSDA-wide file format registry" and "Examine the potential impact on social science research of the expanding number of data file types and information formats, and critically assess the measures that an upgraded CESSDA RI (Research Infrastructure) will need to take in order to adequately acquire, preserve and disseminate these new forms of data".

These two tasks are strongly related. For that reason one combined report is presented here covering the issues related to multiple file formats and offering recommendations for a CESSDA-wide file format registry.

The report starts in chapter 2 with a descriptive overview of expected changes in the use of file formats and tools in the social sciences. This is for a large part based on the survey carried out in 2008 among the CESSDA data archives with some additional answers. What is the background of these changes and what, in particular, is their implication for the long term preservation of data within the CESSDA archives? The consequences of these new trends for the new CESSDA RI will be analysed and discussed. In particular attention is paid to the OAIS reference model and, the concept of "trusted digital repositories" and related guidelines and best practices and its usefulness for the CESSDA RI.

One of the questions asked in this survey was "Please indicate the changes in format that you anticipate and briefly state what plans you have to deal with these". Do the CESSDA organisations actually have plans to deal with the new file formats and if not what should they need to handle these new formats? This is in fact the main issue of this report to be answered in the rest of the report.

In chapter 3 long-term preservation is discussed in a broader context and its main issues looked into. The most pressing problem is that of software obsolescence: the risk that data which is preserved in a certain file format, cannot be read sometime in the future because the software is not available any more. A critical assessment of existing preferred file formats (global format registries) and tools will be given. What are preferred file formats and what is the use of this concept?

Chapter 4 and the appendices, at a more practical level, contain recommendations relating to the construction and maintenance of a CESSDA-wide file format registry and proposal of a preferred formats list.

A list summarising the recommendations of this report can be found at the end.

The report has been written by Data Archiving and Networked Services (DANS) Contributors were Valentijn Gilissen, Henk Koning, Dirk Roorda and Heiko Tjalsma. We gladly acknowledge the valuable feedback received from CESSDA members on the previous version of this report.

2. File formats in the social sciences and their preservation

2.1. File formats in the CESSDA data archives: now and in the future

Social science data archives are the oldest digital archives of the world. From the sixties in the twentieth century on they collected and preserved research data. These data were from the beginning on primarily survey data. That means that they were mostly organised in rectangular files of tabular form and in most cases were analysed with a statistical package program like SPSS, OSIRIS or SAS. (Doorn 2004, 98) In other words: The social science data archives were originally conceived as survey archives. Only much later, from the late 1970s on research data archives for other disciplines came into existence: text archives, historical data archives and archaeological data archives. These archives, organised often apart from the social science data archives, ingested from the start on more varying file formats. In this report we will restrict ourselves to the social science data archives, to which we will refer to from now on as CESSDA data archives. (Doorn and Tjalsma 2007, 3-4).

This historical background of the CESSDA data archives is reflected in the results of the survey carried out among the CESSDA data archives in 2008. It emerges very clearly from that survey that in 2008 most CESSDA data archives still have a very large percentage of data submissions in standard statistical file formats. Only in three out of seventeen CESSDA data archives is, of all the data files ingested, *more* than 20% are *not* in standard statistical file formats (table 1).

Table 1. Approximate proportion of incoming files that are not submitted in standard statistical formats (i.e. SAS, SPSS, NSD-Stat).

Approximate proportion	Number of data archives
5% or less	7
> 5% and ≤ 10%	4
> 10% and ≤ 20%	3
> 20% and ≤ 100%	3

N= 17

Source: CESSDA-PPP Survey 2008; only CESSDA data archives

Most of these “non-standard” statistical file formats are text files and for the rest largely databases and spreadsheets. Numbers of video, audio and image files are comparatively small. It can clearly be concluded that not all CESSDA data archives contain only statistical survey data any more, as they probably all did in the very beginning of their existence. There is a clear tendency towards diversification. This is in line with what we know from other sources: in some data archives other than purely quantitative data are moving in. Some organizations explicitly acquire qualitative data. Within the data archive of the United Kingdom UKDA the well-known unit Qualidata has existed for years and explicitly collects qualitative social science data. In Switzerland FORS also explicitly accepts qualitative data, although it is not their core business and also the Finnish data archive FSD actively acquires these data (Corti 2007, Qualidata, FORS, FSD).

Table 2 Non-standard statistical formats included in CESSDA-PPP organisations.

Formats	Number of data archives
Text files (Word, XML, GML)	13
Spreadsheets, (e.g. Excel)	12
Databases (e.g. MS-Access, Oracle, Filemaker, dBaseV)	9
Audio files	2
Video files	2
Image files (e.g. JPEG, TIFF)	5
Other	3

N = 16, multiple answers possible,
Source: CESSDA PPP Survey 2008; only CESSDA data archives

The CESSDA data archives themselves see this tendency too: in the survey 52.9 % of the archives are expecting that “future changes to the file formats used in the social sciences” will take place. Only 11.8 % do not think so and 35.3 % do not know. When asked what changes are expected then half of the respondents do not know at all and the other half gives widely varying answers, like the “data provider might move away from SPSS or other standard statistical packages” or “SPSS could die out”. Another group of answers highlights, in one way or another, the increase of varying types of file formats, both in the category of statistical data files as well as in the new categories like audio, video, photos and GIS-data.

It is not without reason when it is said that “SPSS could die out”. SPSS originally, to cite from the SPSS website itself, “stood for the Statistical Package for the Social Sciences” and is almost as old as the CESSDA data archives themselves. In many data archives it was for a long time “the” standard and it often still is. It was founded in 1968 by Norman H. Nie, C. Hadlai Hull and Dale H. Bent, who developed a software system which made it possible to use statistics in order to analyze raw data. Its environment was from the start very strongly academic. It was in its early years hosted by the National Opinion Research Center at the University of Chicago. In the years 1975-1984 however SPSS had to separate itself from this Research Center and became a commercial corporation selling commercial software. Since then it has focused itself more and more on the business world, more so, it seems nowadays, than on the poorer and less profitable academic (social science) world. It describes itself at its own website in 2009 as follows: **“**SPSS is recognized as a leader in the predictive analytics market space. Predictive analytics, which combines advanced analytics and decision optimization, will continue to be a focus for the organization as it seeks to increase marketplace understanding of the business benefits that predictive analytics provides.” This shift in orientation of customers together with rising prices for this package and cheaper alternatives are the reason for the widespread expectation that SPSS might not longer be the standard and market leader within the CESSDA community which it used to be for years (SPSS).

In the same way as SPSS could be seen as a de facto standard for software and file formats in the CESSDA data archives, is the DDI (Data Documentation Initiative) the standard for metadata. DDI, which had as its predecessor the SDS (Standard Study Description Scheme), has become an elaborate standard for describing data from social science based on and taking into account the methodology of the social sciences. (Rasmussen and Blank 2004; Blank and Rasmussen 2007)

To sum up: Two different developments are expected by the CESSDA organisations. The one is a change towards qualitative data, like texts in one way or another, video and audio files as well as an increasing use of spreadsheets and databases. The other development foreseen might be a change in the choice of statistical software programs as a standard for using and *storing* the data.

2.2. Multiple formats: consequences for the CESSDA data archives

The observation that many CESSDA data archives expect an increasing variety in the file formats has important consequences. The increase of multiple formats will be felt in a number of activity areas of the data archive, such as staff competencies and training, legal issues, changes in user communities, storage capacity as well as security of the data. We mention these consequences here only briefly and in relation with the topic of this report. They will be dealt with in more detail in other parts of the CESSDA-PPP report. The consequences for data (and metadata) preservation will be dealt with in the next paragraph.

Expansion of file formats means that a detailed knowledge of these new formats is needed by the data archive staff, both data archivists and IT personnel. Better knowledge of some of the old formats would be needed as well. How to handle these data, how to store them and possibly convert them requires new staff competencies. So staff need either to be trained or expanded, with new employees experienced in these new formats. At least a broader knowledge of software systems is needed and possibly more specialisation within the staff. Connected with this is a possible change or broadening of user communities. Qualitative data are data which can also be of great value for historical, linguistic or anthropological research amongst others. One can think here for example of oral history studies. Expansion of user communities means possibly more various ways of communicating with these communities by the data archive.

"The increasing variety among the types of digital records will likely introduce greater complexity to the process of archival reference services, even as technological innovations continue to expand upon the services that are possible", to cite Margaret O'Neill Adams (O'Neill Adams 2007) which is exactly to the point here. She argued that each new type of digital records could involve new models for user services. The first step would be to recognize the differences in the needs and expectations of information seeking requestors (the general public) as compared with requests from persons engaged in original scholarly research (discipline related researchers, also described as the designated community, see paragraph 2.4). In other words: new data types could attract new user communities, dependent on the content of the data and the research discipline but also more diverse use of the data by a general, information seeking, public. A data archive has to be prepared for this.

Not only is new and more widespread knowledge of software (formats) needed, also the knowledge of metadata systems will need extension. The Internet already requires at least knowing how to deal with newer systems as Dublin Core and the Open Archives Initiative OAI-PMH, but descriptive systems might also differ between user communities. Archaeological data for example need different contextual information than social science or linguistic data. The case of oral history studies could be mentioned here as an example. Oral history files could be documented as social science data with the DDI metadata system. They could however also be documented in the way they would be handled in paper archives where (historical) sources are often documented with the General International Standard Archival Description ISAD(G). (Corti 2007)

Regarding legal issues adaptations might be needed as well. This might be necessary in particular when a CESSDA partner should receive datasets containing personal data in any form. Legislation on the permitted use and storage of these data is within the European Union very strict, at least when the persons are identifiable. This could be a new phenomenon for some of those CESSDA archives who until now only ingested survey data which have been made anonymous. When ingesting qualitative data the chances are higher that personal data are included, for example in oral history interviews. Despite the fact that there are directives of the European Union on these issues, legislation is national with mostly minor variations from country to country.

Also on a more technical level adaptations might be needed. Video and audio files require relatively larger amounts of storage than quantitative data. Probably more important is that there might be heavier demands on security as qualitative files can very well contain personal data, even of a sensitive nature, which (still) have to be protected severely.

2.3. Digital preservation in the CESSDA data archives: the present state

A growing number of new and changed formats are expected within the CESSDA data archives. There is no longer the uniformity of the now almost traditional standard statistical packages which used to exist in the CESSDA data archives. The same tendency towards diversity could apply for another standard, the DDI, the standard documentation system used in the CESSDA data archives, in particular when non-quantitative data would be moving in in large numbers. It is still to be seen whether the DDI will remain suitable for all types of non-quantitative data. There are, on the other hand, no clear alternatives at the moment, it seems, so this is speculative. This has consequences for many of the activities of the data archive.

If only a change would take place from one standard into another, from SPSS portable into SAS or STATA for example, then these consequences would be less radical of course. A far more fundamental change would consist of the expansion of file formats as a result of a widening of the acquisition policy of the CESSDA data archives. We assume that this is, or will be, the case in most CESSDA data archives, now or in the near future. Acquisition will not be restricted to the "traditional" surveys, but extended to more varying types of quantitative data (in databases for example) and in particular of qualitative data. Also data from the public administration could be ingested by the data archives in the new CESSDA RI, acquired either from the national statistical institutes or directly from government departments or agencies.

This expansion of file formats used has significant consequences for the topic to which this report is devoted: the long term digital preservation of the data. The fact that statistical packages are no longer the only file format around is important for the digital preservation of the data.

Table 3. Standard preservation formats used in CESSDA data archives.

	Formats	Number of data archives
1.	SPSS, SAS, STATA, NSDSTAT, OSIRIS	11
2.	CSV OR ASCII FORMAT	7
3.	XML, SGML	3
4.	PDF	2
5.	DATABASE, MYSQL, EXCEL	2
6.	RTF	2
7.	TIFF	2
8.	HTML	1

N = 17, multiple answers possible
Source: CESSDA-PPP Survey 2008; only CESSDA data archives

When asked on which file formats the CESSDA data archives rely for digital preservation the statistical packages are clearly mentioned most times (table 3). Some data archives have explicitly stated this in additional comments as well. This is, historically seen, understandable as these packages often were the only ones used in the data archive. From the results in table 3 it can be derived that the statistical software packages are the single most used software format, in particular SPSS[1], for preservation at the CESSDA data archives followed by "plain" CSV and ASCII formats. Combining these findings with the ones in table 1 (paragraph 2.1) on incoming files it seems safe to assume that the statistical, quantitative, data are contained, and in most archives probably remain contained, in the statistical software packages or that they are stored as plain text data (ASCII etc.).

The most commonly used format is probably the SPSS portable format which is very well suited for transporting SPSS system files across different platforms (UNIX, Windows, MVS etc.) as well as across different versions. It should however be realised that this format was

[1] According to the survey SPSS files were either saved as system file (.sav) or portable file (.por). In the majority of answers the exact SPSS format was not specified.

never intended or constructed for preservation aims (see for more on this in chapter 4 of this report)! The SPSS portable format is what it says: to transport the data, not necessarily to preserve them. The SPSS system file format (.sav) is even less suitable for preservation aims.

Combining this knowledge on the actual use of standard preservation formats with our earlier observation on the increasing variety of file formats gives reason for concern. How well spread is the general knowledge or even awareness amongst the CESSDA data archives on the subject of digital preservation?

In the survey amongst CESSDA data archives it was asked whether the data archives have a "preservation policy" and if that would not be the case to outline the basics of a de facto preservation policy. It should be realised here that preservation policy is of course a broad term covering both preservation of hardware, software, file formats as well as metadata (see also chapter 3). A variety of answers was given. A minority of CESSDA data archives actually has a preservation policy covering all these aspects. By a number of data archives the question was answered by referring to media renewal and back-up systems, without mentioning what to do with the possibility of software becoming obsolete. Answers like “datasets are maintained in their original formats” and “quantitative data are preserved in SPSS portable format” clearly indicate a lack of awareness of the digital preservation issues.

The answers to the question “Which actions do you take to ensure the long-term preservation of non-standard data files (i.e. not SPSS/SAS/NSD-Stat)” are also not reassuring. The two most common answers are “Migrate them into one or more standard archival format(s)” and “Store them in their original format” (table 4).

Table 4. Actions taken to ensure the long-term preservation of non-standard data files (i.e. not SPSS/SAS/NSD-Stat)

Actions	Number of data archives
Migrate them into one or more standard archival format(s)	13
Store them in their original format	8
Otherwise convert them	3
Other*	2

N= 16, multiple answers possible

Source: CESSDA-PPP Survey 2008; only CESSDA data archives

*Other:

- Imported to a database
- None as of this moment
- NESSTAR, NSDstat, DDI, XML
- Usually relying on data provider to take primary responsibility for archiving

Long-term preservation of data files consists in most CESSDA data archives of migrating to standard formats (mostly the statistical packages or text files), to summarize the findings of the survey. This is in itself not necessarily a bad thing, but it really is not certain that the formats used will prove to be the best for long term preservation. Furthermore the results of the survey cast doubt on the level of knowledge and awareness on this topic within the CESSDA data archives. The low number of archives having a preservation policy is not satisfactory either. Maybe things are not as bad as it seems, but all together the situation regarding digital preservation can be described as worrying.

2.4. The use of OAIS, trusted digital repositories and best practices and guidelines for digital preservation in the CESSDA RI

The worrying conclusion in the preceding paragraph on the state of digital preservation within the CESSDA community as a whole leads us towards the question what would be needed to reach a standard high enough for the CESSDA RI. To answer this question, in the form of recommendations, we will present in the next two chapters a broad overview of the issues in digital preservation, in particular preferred file formats.

Before we do that we need to pay attention to the concepts of OAIS (Open Archival Information System) model and of "trusted digital repositories". These are both, in a different way, important general models for achieving quality in digital preservation. They are both increasingly used by all those organisations that need to preserve digital material of any kind. As they are general models they however require further elaboration in the form of specifications, guidelines and best practices. These are very much under construction at the moment.

OAIS trusted digital repositories and related guidelines and best practices are of great importance for the construction of the CESSDA RI. This is clearly indicated by one of the proposed *Obligations of Full Members* of the CESSDA RI: "To adhere to the principles of the OAIS reference model and any agreed cessda-ERIC requirements for operating trusted data repositories (The CESSDA-ERIC 'seal of approval')" (Privileges and Obligations of CESSDA RI Membership, draft version, 1.2 h).

In the remaining paragraphs of this chapter we will present a short introduction into these models and related guidelines and a recent evaluation of their applicability for the CESSDA RI. For a more detailed overview we refer to the report "Recommendations concerning best practices", task 6.4 report of the CESSDA-PPP (Štebe and Dusa 2009, Fábián 2009).

2.4.1. The OAIS model

According to the model an OAIS should primarily be seen as an archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a *designated community*. The *OAIS reference model* is defined by a recommendation of the Consultative Committee for Space Data Systems, as its origins lies with NASA (OAIS Blue Book 2002) [2]. The model can be used for archiving all kinds of electronic material, ranging from electronic publications to data or other digital objects. Because of its general nature it can be applied to a wide range of institutes carrying out preservation tasks of digital material. It contains all the processes needed but only in the form of a framework. The OAIS itself as a reference model does not proscribe standards, guidelines or best practices. Further specifications are necessary to enable the execution of the many tasks in archiving digital objects for data archives, repositories, libraries and public record offices. Orientation towards the OAIS is taking place in a growing number of these institutes.

Four major elements can be distinguished in the model

- A vocabulary enabling communication on common operations, services, and information structures of repositories;
- A simple data model for the information that a repository takes in (or "ingests", to use the OAIS vocabulary), manages internally, and provides to others;
- A set of required responsibilities of the repository for negotiations with producers of information to get appropriate content and contextual information;
- A set of recommended functions for carrying out the archive's required responsibilities. These are broken up into six functional modules: ingest, data management, archival storage, access, administration, and preservation planning. (Ockerbloom 2008)

An important notion within the OAIS model is the distinction in a designated community ("primary users") and secondary users (OASIS Blue Book 2002, pages 1-11). It means that the OAIS is basically intended to work for and with a designated community of consumers to make sure they can independently understand this information, and follow well-defined and well-documented procedures for obtaining, preserving, authenticating, and providing this information (Ockerbloom 2008).

There are, however, also other user groups, who might use the data, ultimately the general public. According to the model an archive may decide that certain content information should be understandable to the general public and, therefore, by broadening the definition this becomes the designated community. For example, information originally intended to be understandable to a particular scientific community may need to be made understandable to the general public of the designated community (OASIS Blue Book 2002, pages 3-3 and 3-4).

The OAIS offers an organisational and functional framework and is formulated in general terms which makes it flexible for different kinds of organisations (Fábián 2009). It should be realised that because of this the OAIS model can function very well as an abstract framework for *communication* about the long-term preservation of digital materials for a specified designated community. To illustrate this we refer to one of the conclusions of a report

[2] There is a draft version of a new version of OAIS containing improvements which will be submitted to ISO for review: http://cwe.ccsds.org/moims/docs/MOIMS-DAI/Draft%20Documents/OAIS-candidate-V2-markup.pdf

published a few years ago which said that the use of the OAIS as such is very helpful as a means of communication between two (or more) archives operating in quite different settings and organisational structures. This was in a study aimed at finding out how comparably "OAIS-compliant" the UK Data Archive (UKDA) and the British National Archives were. The conclusion was that they were certainly compliant in broad lines but that it is very time-consuming to map all the details of the functional model (Beedham 2005, pages 4-8, 81-84)

Vardigan and Whiteman concluded in an article in 2007 in which they evaluated a mapping exercise of the ICPSR to OAIS that it “ultimately” is possible to design a federated system of trusted social science repositories: “Conforming to the OAIS standard will permit the archives to communicate more effectively and to provide access to a network of trusted digital repositories.” (Vardigan and Whiteman 2007)

According to the UK Data Archive study one of the important points in the OAIS model was "the strong link between the user community and the way the material in the archive should be described and preserved". It is difficult to limit user groups or communities as narrow as the OAIS does in its concept of designated community. The UKDA and the National Archives were simply not able to identify clearly described and homogeneous user communities. This is the same point raised earlier by Margaret Adams as quoted in paragraph 2.2 on the widening of user groups.

Not only in this UKDA report, but also by others it is often seen as a disadvantage that the OAIS documentation is quite long[3] and complex and that this may prove to be a barrier to smaller repositories or archives (JISC Standards Catalogue and Beedham 2005, page 82). Another point is that pre-ingest is a subject which is hardly covered in the OAIS. Negotiations with producers of information to get appropriate content and contextual information are part of the OAIS, but besides that the model does not cover what to do *before* the data actually have arrived at the data archive. The whole process of acquiring data is neglected. We will evaluate these critical comments in paragraph 2.6.

2.4.2. Trusted digital repositories and the use of best practices and guidelines

As “trusted digital repository” is a broad concept. Consequently these guidelines cover many different aspects of digital preservation, varying as they may be in scale or approach. We have briefly discussed most of these aspects earlier in paragraph 2.2. They include data preservation, staff competencies, legal issues, storage and security. In this report we will not go into the details of all these aspects with the exception of data (and metadata) preservation. This issue is of course the most essential one in the best practices and guidelines on trusted repositories.

[3] The OAIS Blue Book containing the model of OAIS is 148 pages long.

Trusted digital repository as a concept implies that data will be preserved in a safe place so that they can be accessed in the same way now as well as in the future. . Therefore it is basically just as concerned with digital preservation as the OAIS does. This concept is extremely relevant in the context of digital preservation of research data and therefore also of great importance for the construction of the CESSDA Research Infrastructure. An additional stimulus for developing the idea of trusted digital repositories is the growing tendency to store data increasingly in local university or institutional repositories, different from the, often more centralised, national data archives or electronic deposit libraries.

A trusted digital repository should have incorporated in its regulations the whole relevant legal framework needed for digital preservation. This means that the national legislation should be followed not only concerning the protection of personal data (paragraph 2.2), but also that on intellectual property rights. In the new CESSDA RI licence contracts as well as user conditions should be made up-to-date with these rights. This should include attention for the issue of the legality of copying data files for preservation purposes (migration or conversion) by the data archive. On this the legal situation is quite different within the European Union. In a number of countries now academic "codes of conduct" for the exchange of knowledge and information exist, either of a general character or specifically aimed at using personal data in academic research. These should be included in the user conditions as researchers have to agree to these codes when using data.

In this field orientation towards the newest developments in the Creative Commons Movement would be strongly advisable as well. For data a Creative Commons licence model is not yet available, contrary to the one for electronic publications, but this might come in the next years. Another important movement in this respect is the Open Access Movement, aimed at an open and free distribution of academic publications as well as research data. (Creative Commons, Open Access, see also Štebe and Dusa 2009, section 2.1).

Like the OAIS it is a general concept, so, again, a set of criteria needs to be formulated to ensure that a trusted digital repository meets minimum standards of quality, traceability, accessibility and usability. Ultimately this should lead to criteria and procedures for the *certification* of digital repositories. A number of guidelines is now being developed by various organisations trying to formulate guidelines and/or best practices for "trusted digital repositories". These guidelines could also be described as Data Activities Reference Models (DARMs), (Štebe and Dusa 2009). We restrict ourselves to the most relevant or promising ones here.

- TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist), preceded by the *Audit Checklist for the Certification of Trusted Digital Repositories,* was originally developed by RLG and NARA (Research Libraries Group and National Archives and Records Administration). It contains criteria and procedures for the certification of digital repositories and is basically an audit checklist. It distinguishes sections on organisational infrastructure, digital object management and technical infrastructure. Efforts are now under way to evolve TRAC into ISO standardisation (see http://www.digitalrepositoryauditandcertification.org/).

http://www.oclc.org/programs/about/collaborations.htm
http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91 (Ruusalepp 2009).

- DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) published by *DPE Digital* P*reservation Europe/DCC The Digital Curation Centre presents itself as a* toolkit for a Risk Management Model. The aim of the model is to help organisations to develop an organisational profile and, in particular, to identify and assess the risks "that impede their activities and threaten their assets" (Ruusalepp 2009).
- In Germany the digital preservation initiative NESTOR (NEtwork of Expertise in Long-term STOrage of Digital Resources) is a catalogue of criteria for trusted repositories. It distinguishes three major areas: organisational framework, object management and infrastructure and security. In spring 2008, NESTOR handed over its standardisation projects to the German Institute for Standardisation (DIN). (Schumann 2009).
  These are all fairly detailed guidelines; The DSA Data seal of Approval, developed by DANS in the Netherlands, contains on the contrary broadly formulated guidelines intended to ensure that also in the future research data still can be accessed in a reliable manner without requiring the implementation of new standards, regulations or high costs. As the guidelines of the seal approval are not worked out at a detailed level the consequence of this approach is that trust plays an important role. Preferred formats are an important element of the DSA. (Data Seal of Approval)

For an evaluation see the next two paragraphs.

2.5. Trusted digital repositories and the CESSDA data archives: the The Hague workshop

In January 2009 a workshop was organised in The Hague (the Netherlands) by DANS and UKDA within the framework of work package 10 of the CESSDA-PPP project. It was aimed at the applicability of these guidelines in the social sciences. The main issue was: do the now existing guidelines fulfil the need to accomplish trusted digital archives for the social sciences, in particular for CESSDA? Special attention will be paid here to the conclusions of that workshop as they have a high relevance in the setting of this report (CESSDA Digital Preservation report 2009)

The conclusion of this workshop for the CESSDA participants was that as a first step the DSA guidelines could be useful to create awareness and to set digital preservation, or permanent access, in motion. The DSA guidelines are formulated in a broad and general way and are not only on data repositories but also on data producers and data consumers. When on an operational level more precision is needed, in particular for professional organisations, they are however not enough for assessing and evaluating repositories. Methods of a detailed level such as TRAC and DRAMBORA could be useful as additional steps towards certification, but when it comes to using these two checklists in practice they are very elaborate and expensive.

In the DSA self assessment is foreseen as general assessment procedure. This method was seen as a crucial point. For an organisation like CESSDA where mutual understanding and trust is high a self assessment method could be acceptable. For other organisations and groups this could be quite different, especially for those who have become acquainted with each other only recently. At the moment the criteria for being acknowledged as a digital repository are still unclear. How rigorously will the DSA seal be requested from newcomers? How rigorously will the assessment be executed? Which peers will do the assessment, how familiar are they with the assessed institutions?

Another point is how a repository could control all its depositors, especially researchers, in short the data producers? Is a repository going to tell a data producer what to do and what not to do? Another missing point is protection of confidential data, especially personal data, outside but as well inside the repository. The interests of the data subjects (see Matthew Woollard's presentation at the workshop Woollard 2009) should be included in the DSA as well. What is also missing in the DSA is the perspective on long term succession planning, the durability of repositories. How long will the organisations stay alive? Are there guarantees when repositories should disappear?

When looking at the possible structure of the CESSDA RI the DSA could be part of the mechanism for entrance of new service providers in the CESSDA legal entity. These new service providers will have to be evaluated and accredited as trusted digital repositories. There should be a threshold to pass, but a self assessed DSA would certainly not be enough for that. For CESSDA continued *access* to research data could be the main criterion for admission, implying of course taking good care of the preservation of the data. Storing and preserving of the data could be outsourced, as some CESSDA partners do now, but the responsibility remains always with the CESSDA partner itself

As we observed before, only a minority of the CESSDA data archives actually has a preservation policy covering all the aspects of digital preservation. The number of data archives which can be considered as being a full trusted digital repository applying all the available guidelines and best practices is at the moment probably still very limited.

2.6. Evaluation of OAIS and trusted digital repositories

2.6.1. OAIS

Using the OAIS could offer the CESSDA RI great advantages. Being a reference model is possibly the strongest point of OAIS, especially seen from the perspective of a CESSDA RI. As such it gives a framework and ontology for communication. It also leaves room for various practical applications tailored to the specific organisational context of a certain repository. The proposed *Obligations of Full Members* of the CESSDA RI restrict themselves to stating that partners should adhere to the *principles* of the OAIS reference model. Other promising points, to be developed further, could be the usefulness of OAIS as a reference framework for calculating and comparing archiving costs as well as staff positions (Štebe and Dusa 2009, Beagrie e.a. 2008).

The disadvantages of OAIS for smaller institutions should not be that much of a problem when a CESSDA-wide organisation would come into existence. We would recommend that the CESSDA RI develops its own version of guidelines or adheres to other, still to be developed, guidelines. It should be easy to create a reduced version of these guidelines, adapted for small-scale institutes (as already suggested in the UKDA report of 2005 (Beedham 2005, page 82). The CESSDA community (the CESSDA archives and users taken together) is at the moment of course a clearly designated community which could also be helpful. As said before (paragraph 2.2) it should however be realised that this designated community will certainly widen in the near future which is a point of attention. This is in line with the conclusion of UKDA in 2005: it will be less easy to discern user communities.

In the CESSDA RI where national organisations often work in varying national frameworks and conditions it does not seem to be a problem that the pre-ingest phase is hardly worked out. Pre-ingest could become an explicit part of the CESSDA RI organisational structure, but it should then be elaborated in the accompanying guidelines.

2.6.2. Trusted digital repositories

There is no universally accepted standard yet leading up to certification of trusted digital repositories. The guidelines and best practices trying to work out this concept at a practical level operate at quite different levels. Some are very detailed audit checklists or methods focusing on the organisational and technical infrastructure (TRAC, DRAMBORA), others are on the other hand very general and not very specific or detailed (DSA). Generally speaking, one of the difficulties is that, if guidelines contain technical specifications, they may be outdated soon and have to be updated regularly, because of rapid technological developments.

According to *The Task Force on Archiving of Digital Information* in 1996 “a process for certification of digital archives is needed to create an overall climate of trust about the prospects of preserving digital information.” (Preserving Digital Information 1996). Trust is indeed the key word here. That makes it a central issue in all these different guidelines, the basic question being how reliable a trusted digital repository really is. Trust should be an essential necessity within the CESSDA RI, between the users and the data archives as well as between the data archives mutually. This should be the decisive element in formulating or choosing guidelines and best practices on trusted digital repositories.

2.7. Conclusion

An increase of qualitative data, i.e. multiple file formats, and statistical software programs losing its position as a standard for using and storing data are to be expected in the CESSDA data archives in the coming years. Combining these observations with our conclusion on the present state of digital preservation gives reason for concern.

Guideline 7 of the Data Seal of Approval (DSA) prescribes: "The data repository has a plan for long-term preservation of its digital assets". At the moment only a very limited number of CESSDA data archives have such a plan. Using the OAIS and the trusted digital repository models will greatly add to establishing the quality of the digital preservation in a CESSDA data archive. Relations within the CESSDA RI should be based on mutual trust.

Our recommendation would be that, to make this latter conclusion work for the future CESSDA RI, two conditions should be fulfilled. The one condition is the construction of a clear set of guidelines, as formulated in the proposed *Obligations of Full Members of the CESSDA RI*: "any agreed CESSDA-ERIC requirements for operating trusted data repositories". This would imply a kind of cessda-ERIC 'seal of approval' which could be the DSA supplemented with solutions for the issues mentioned in this workshop.

The other prerequisite is that for the CESSDA RI assessments are needed of each partner to make sure these requirements, in the form of guidelines, are met. The assessments should be appropriate to the level of service which the partner is going to provide.

3. Overview of long-term preservation issues & critical assessment

Digital preservation is a problem that is easy to understand. It is harder to identify it at various levels in digital information management. And it is difficult to employ concrete measures to guarantee the sustainability of digital information.

The effort to define preferred formats for research data is such a measure, and a valuable one, because, when nothing else is done, insisting on preferred formats is a huge improvement in the preservation of digital data.

Not all facets of digital preservation are covered by file formats. It might help readers to get a quick overview over the digital preservation landscape first. Subsequently we focus on the aspects touched by the issue of preferred formats.

3.1. Fundamentals of long-term preservation

Long-term preservation of digital data denotes all the care, measures and activities to ensure that the data at hand will be still usable at future times either by the designated community or the general public, even when the present systems of digital information processing have become obsolete.

There is the hidden assumption that there will always be information processing systems capable of dealing with information of the same complexity as we do now. With this assumption, the problem is basically one of backwards compatibility: what do we have to do to ensure the correct processing of information in newer and usually more powerful systems?

In order to assess the problems of long-term preservation, it is helpful to divide the phenomenon of digital information into three layers: (a) hardware, (b) software, and (c) human knowledge.

Layer (a) contains the mechanisms in which the digital ones and zeroes are realised in the physical parts of reality, such as electronics, magnetism and optics. Layer (b) contains the technology of coding meaningful information into bits and bytes, and processing the resulting digital data. Layer (c) contains the background knowledge that human producers and users of information employ to make sense of the data, to assess its importance and to use it as basis for new information.

There is a hierarchy in these layers: without any hardware, there will not be a software representation, and without the latter, there will be no human knowledge. However, there is also a fair amount of simplification here, because in computer science the boundary between hard and software tends to become blurred. Yet this perspective of a threefold division between data in the middle, its sustaining technology at the bottom, and its interpretation by humans on top, is a fruitful one. It is not obliterated by the nuances of computer science; rather it repeats itself in various contexts. For example, programmers developing SPSS software think statistical algorithms (top level), produce program code (middle level), which is translated by a compiler into machine code. This is the SPSS developer's perspective. But from the outside we see that all these levels are in fact purely software.

From this a useful tenet in digital preservation can be distilled: in order to preserve digital data with any level of confidence, it is necessary to take the sustaining technology into account as well as its interpretation by humans.

3.1.1. Problems

The problem of long-term digital preservation is that over time there occur a lot of changes in all three layers of a given set of information.

(a) The hardware layer of data, the physical representation of a dataset is subject to the laws and contingencies of nature, including the tendency to become progressively disorganised. This problem is called the problem of *bit stream preservation*, or: how to guard against *bit rot*. Examples are: hard disk, floppy disk, CD/DVD, tape, internal memory. Here the word bit stream refers to bits in layer (a), the physical representation, rather than to layer (b).

(b) Quite a different kind of change is happening at the software level. The whole machinery of processing digital data is a complex field of operating systems, low- and high-level computer languages, network protocols, and utility programs. The organisation of this field is the direct product of an explosive growth in underlying hardware capabilities and an accompanying optimisation and integration of information processing programs. In this rapid growth there is only a limited degree of backward compatibility. The general term by which this problem is known is: *software obsolescence*. Here software stands for all products of information technology, such as programs, file formats, protocols.

(c) The evolution of human knowledge, which might even be called an explosion, presents a third difficulty. Old methods and background knowledge and hidden assumptions will be forgotten; it will be hard for posterity to understand our results and data in the form that we leave behind, unless we take precautions. In punched cards, for example, multiple punches may more easily be understood by those who understand a base 12 system (like those who used a non-decimal currency). The problem is to preserve the *intelligibility* or *interpretability* of the information. Another problem is reference to external information. Whether or not such references remain valid, is a matter of human organization.

3.1.2. Solutions

From the description of the problems in the layers (a), (b) and (c) it will be clear that the solutions will be vastly different. They are to be found in different disciplines, and to be provided by different competences.

(a) There are two main ways to improve bit stream preservation.

The first one is to use reliable media, that do not decay easily, and for which the read/write processes are robust. The media should then be stored safely, protected against disasters and intentional damage and replaced at regular intervals. This kind of reliability and safety needs technical and organisational development.

The second method is a hallmark of digital data: redundancy. Digital data is easy to copy, and the threats to physical bit streams are unlikely to threaten all copies at the same time.

Concerning the first method: technology is in the lead here. The process of condensing storage capacity, speed up access times, improve reliability and reduce cost is very much going on, with periods of marginal improvements followed by explosive bursts. Currently the hard disk is the storage method of choice for large scale storage of accessible information. But the hard disk has a hard time to keep up with the speed of modern processors, so there is intense research for alternatives, such as solid state disks or more exotic technologies. The information curators are not the ones to steer the technological development. Their responsibility is to find appropriate ways of organizing digital data in such a way that the risk of data loss is balanced against the costs of preservation measures. Most of this organisation involves redundancy and mixing solutions (using both magnetic drives and optical media, for example).

Concerning the second method: managing multiple copies is a challenge for information curators. On the one hand you need software to take care of automatic synchronisation of copies to changing originals, but on the other hand you do not want to lose copies automatically when the original gets lost accidentally. This approach calls for careful identity and restore management. See the publications of the LOCKSS (Lots of Copies Keep Stuff Safe) project (LOCKSS, Chronopolis)

(b) There are also two main methods to combat software obsolescence.

The first one is to preserve the original software environment. There are two flavours: using original equipment or emulating the original environment on new technology. Since information technology does not hinge on specific hardware, emulation is much more natural. Software is not defined by its piece of hardware, but by its interface in terms of bits, algorithms and protocols, which can be implemented on any hardware that supports generic computing. We shall refer to this method as *emulation*.

The second method accepts the ongoing change, and adapts the data to new environments. The term by which this method is known is *migration*.

There are subtle and less subtle pros and cons to emulation and migration that we discuss shortly. The upshot is that both are important, that they are not completely independent, and that the nature of what is to be preserved is a decisive factor when it comes to choosing between the two strategies.

See on emulation:

- The Rand report “Addressing the uncertain future of preserving the past, Towards a robust strategy for digital archiving and preservation” (http://www.rand.org/pubs/technical_reports/TR510/)
- the MIXED (Migration to Intermediate Xml for Electronic Data) project (http://mixed.dans.knaw.nl/node/114)

- Digital Document Quarterly (DDQ) (http://home.pacbell.net/hgladney/ddq.htm) and Preserving Digital Information, (http://home.pacbell.net/hgladney/hmgpubs.htm#book)
- The KEEP (Keeping Emulation Environments Portable) project (http://www.keep-project.eu/ezpub2/index.php)

(c) There is no method that guarantees the preservation of intelligibility of information. We do not know how to define what we mean by (human) intelligibility in such a way that we can prove that we preserve it. But we only have to look at the disciplines of history, linguistics, anthropology, hermeneutics, in order to see that we can understand a lot of what has been handed over to us. The other insight is that it also costs an inappropriate amount of resources and ingenuity to do so, and that a bit of extra information can reduce the effort dramatically, the Rosetta Stone being the prime example here.

Instead of dealing with this problem in a fundamental way, it is better to deal with it in an economic way: what pieces of information can we add to facilitate posterior understanding? So the method here is *metadata*, and the issue is: what is the maximum amount of metadata that the producer of information can be bothered to provide, and what is the minimum of information that contributes to future understanding? See for example the section on Information Packages in the OAIS model (OAIS Blue Book 2002)

The stronger a methodology in a discipline is, in terms of adherence to explicit criteria, the easier it is to guard against loss of interpretability. Provided, that is, that the methodology itself is documented well. Not every dataset has to document the complete methodology, but there should be preserved sources of encyclopaedic information that can be consulted, and it would be very convenient indeed if datasets contained pointers to such information.

Where there is a lack of established methodologies, there should be enough descriptive metadata attached to datasets to make them “independently understandable”, such that users of that data from outside the original context are able to make sense of it. What “enough” means, is a not so easy question. Facing the fact that it cannot be answered exhaustively, we need updatable metadata systems, where future users can add their interpretations of the data. When established methodologies exist the material should be understood by an end-user whose training includes basic research methodology in that discipline.

As to the problem of the validity of references to external sources: a promising solution is the use of persistent identifiers. These are abstract codes that identify a resource, without locating it. An associated service, called a resolver, translates a persistent identifier into one or more valid locations. The burden of keeping references valid now rests with the maintainer of the tables that underlie the resolver service. This requires good organisation, nevertheless this method is much easier than maintaining the references themselves. See for example for plans for this: http://www.icpsr.umich.edu/cocoon/ICPSR/FAQ/0250.xml or http://www.icsti.org/documents/PressReleaseMarch2009-JointDOIforData.pdf

Some other relevant links are:

- The CASPAR (CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) project (http://www.casparpreserves.eu/) and more particular the report Prototype of Descriptive Information-related KM (Knowledge Management) Services. (http://www.casparpreserves.eu/Members/cclrc/Deliverables/prototype-of-descriptive-information-related-km-services-1/at_download/file)
- The DDI (Data Documentation Initiative) project (http://www.ddialliance.org/) and the DDI Strategic Plan 2007 and Beyond (http://www.icpsr.umich.edu/DDI/org/strategic-plan.pdf)
- The Rosetta Stone, a valuable key to the decipherment of hieroglyphs (http://www.britishmuseum.org/explore/highlights/highlight_objects/aes/t/the_rosetta_stone.aspx)

3.1.3. Focus

For the discussion of preferred file formats, the problems in the software layer (b) are most relevant. These problems are on the agenda of users of software applications, their vendors, standard bodies and professional associations around the various scholarly disciplines. Especially in research infrastructures, the way these problems are dealt with becomes a defining characteristic of the infrastructure.

The problems in the layer of the semantics of information (c) are also relevant, because when information is reused in research, it must be very clear what the meaning and reliability of that information is. The fixing of required metadata is a defining characteristic of research infrastructures as well. Usually, there is a trade of between metadata that is stored with the data, and metadata that lays further back: documentation of standards used, prevailing thesauri and ontology’s in the disciplines, reference works and hand books, examples and unsolved problems. As far as attached metadata is concerned, the issue of metadata is a relevant one for preferred formats. But we will not say much about the “stand-off” metadata, i.e. the metadata that is not attached directly to the data, in this chapter.

3.2. Software obsolescence and file formats

Having sketched the fundamental landscape of long-term digital preservation, we turn to the concrete issue of software obsolescence and file formats, an issue that is firmly anchored to the software layer.

In the digital world, data and programs go hand in hand. A tenet in computer science is that the shape of data is determined by the operations you need to perform on them. Although a sound approach from a purely technical viewpoint, this paradigm has had consequences that we just have started to recover from.

As applications change because of user-demands, the form of the data that applications handle has changed accordingly. Aspects of the handling of data have become mixed up with the primary meaning of data. This causes problems in interoperability of data across applications by different vendors and applications across time, even from the same vendor. The problem of interoperability across vendors has been mitigated by the fact that some vendors came to dominate the world around kinds of data, and the problem of interoperability across time has been worked around by a mixture between backward compatibility and migration provided by the vendor.

A good example of this development is what happened with the Microsoft Office formats. In the 1990's we have seen the upsurge of Microsoft Word at the expense of rival WordPerfect The interoperability problems between Word and WordPerfect decreased because the overwhelming majority of users shifted, voluntarily or involuntarily, to Word, and Microsoft added fairly good WordPerfect converters to Word.

In this way, the dynamics of developing new tools, new visualisations, new aggregations of data, has come into the hands of monopolistic vendors, and the long-term preservation of data has become dependent on those same vendors. This is called *vendor lock-in*, and today governments and publicly paid institutions are increasingly aware that *vendor lock-in* is unwanted and should be pushed back.

Continuing our example: Microsoft Word at around 2000 was a good example of vendor lock-in. There were already emerging open formats for text processors, but in practice you needed Word all the time in order to be interoperable with business partners, academic colleagues and government bodies.

Some vendors have reacted to these developments by opening their formats to the scrutiny of the public and tool developers, by pushing their formats forward as open standards. Moreover, they are increasingly aware that their formats are an important factor in the long-term preservation of the data that their applications have helped to create. The new open standards move away from the direct correspondence with the traits of particular applications, and the mixing between application and primary semantics is decreasing, or at least made explicit.

It is interesting to look at the evolution of formats that Microsoft Word can handle. At some point in time it could import and export HTML. From 2003 it had an XML interchange format, with lots of application specific mark-up tags. But the newest 2007 versions of Office use an open, standardized XML format as their native format, doing away with the binary formats and with the pseudo binary format RTF altogether. This also holds for other MS products like Excel, PowerPoint and Access.

This *good trend* is reinforced by the fact that some new open standards have adopted XML as their vehicle of expression. There is a twofold bonus in this: (i) XML is an excellent way to express the structure of data and (ii) XML forces UNICODE as character representation, so that the problems with the inconsistent representation of most of the characters of humanity have become something of the past. We must add here that at the level of compound data there still exists a lot of ambiguity, the perennial problem that most computer systems are preconfigured to deal with dates in a US format!

Despite these positive trends, the problem of obsolete file formats has not yet disappeared. In the first place there is an enormous amount of legacy material in old application formats, and some of these formats are still being used for new material. And, secondly, the move from application formats to file formats has just started and is by no means completed. These formats still contain a lot of application details, which just have been translated to XML. What is still lacking is the notion of a *preservation format*, which is a format that contains a careful selected set of features that are relevant for preservation, and no other features.

Looking at our example for the last time: the Microsoft Office Open XML format (http://en.wikipedia.org/wiki/Open_XML) has a specification of almost 7.000 pages. This is partly because the format should be able to support all features of all previous versions of Microsoft Office. The rival standard, Open Document Format (http://en.wikipedia.org/wiki/OpenDocument), as used by OpenOffice, is much more compact, and is using a quite different approach of text mark-up. As yet, no single format for preservation has evolved for office documents.

3.2.1. Migration versus emulation

Two methods for solving software obsolescence have been mentioned: migration and emulation. Migration is the strategy to convert data from old formats to new formats, preserving meaning. Emulation is the strategy to preserve the original environment of the data: the operating systems and utility programs that were used in creating the data.

Migration is quite an effort and a seemingly endless one as well. An archive needs to be permanently aware which of its holdings are in formats that are currently going out of use. Once the need for a migration is detected, it requires special effort to find or create the software needed for the conversion. After the migration, the new data must be checked for correctness. Finally, the new data should be stored, linked to the original data, with updated provenance metadata, and the administration for the obsolescence detecting facility (whether automated or human) should be updated. This effort can be shared by many organisations which each develop migration tools for a certain type of data. Examples are the DExT project for SPSS statistical files (http://www.jisc.ac.uk/whatwedo/programmes/reppres/dext.aspx) and the MIXED project for databases (as a start).

Emulation is not without pain either. Usually, there is quite a bit of supporting software that an application uses. Everything must be right: the version of the operating systems, the particular configuration and patch level of the web browser, the network and printer drivers. All these elements are updated with uncorrelated frequencies. It needs a lot of administration to be able to run in 2059 a certain application in a particular environment of 2009.

Migration is strong in that it allows forgetting inessential details of the past. This is important for research data, because the purpose of preservation is not to maintain the original look and feel of the computing environment, but to get access to the data, and operate on that data with new analysis tools, or aggregate that data with data from other times and places. For the purposes of a museum migration might be less to the point, though.

Emulation is strong in that it preserves the data as a holistic experience. It preserves look and feel and action (possibilities of data manipulation, performing capabilities for audiovisual content, games). For research data these are often things to abstract away from, but not always. For cultural heritage institutions emulation might be something they cannot do without.

At the end of the day, migration and emulation need each other. File formats have become notoriously complex, and usually the explicit documentation leaves freedom to applications for interpretation. So the final authority on some details of a file format lies in the applications that use them. It is a tremendous help for developers to have access to running versions of these applications while coding format conversions.

Conversely, even when you have complete access to old data through emulation, there comes a point that you want to use that data outside the preserved, emulated environment. That is when migration comes in after all.

The upshot is that for most of the scientific data in the social sciences, migration is preferred above emulation. It pays, therefore, to look for ways to do migration efficiently.

3.2.2. Smart migration

Above we signalled that migration is an endless task, because file formats come in an unending sequence of new versions. Here are a few indications that in reality there is a bit more structure that can be of help. In the first place, we repeat the observation made before, that file formats are just beginning to be decoupled from application formats. They become standardized as well. That means that their update frequency is lowering, which is good news for the migration strategy. Secondly, these new file formats are increasingly open and expressed in XML, like the Open Document formats. That means that their intelligibility is guaranteed, even if there is no longer an application that interprets that format. It will not be too difficult to write a new application that again interprets that format and shows the data coded in it. Essentially that means that the chain of migrations stops there. Of course, even XML will become obsolete, sooner or later (but it will remain “understandable”). But even then we will not have to migrate from XML in order to retain the interpretability of the data. We may have to migrate in order to improve their usability, though. But the sting of data loss by software obsolescence has been removed.

Smart migration is the practice of migrating vendor-specific application formats to application-neutral, standard, XML formats. This migration should be done as soon as possible, when a dataset is ingested into an archive. The data can lie dormant in the archive for ages, until, in the course of time, they will be converted to the application formats of the time, for dissemination purposes. These conversions will be done on the fly, and can be programmed with relatively low effort (see DEXT and MIXED).

The main prerequisites for smart migration are these open, standardised file formats in XML. Not every kind of data that is of interest to the social sciences already has such formats. The existing formats comply in different degrees with the requirements of openness, standardisation, and XML expression. Here is why there is scope for archives to express their preference in the choice of a file format.

3.3. Definition of Preferred formats

Researchers (data producers and data consumers) and archivists (data custodians) are stakeholders with not necessarily the same interests. When it comes to file formats, researchers want to be able to use cutting-edge applications to create and handle their data, and archivists want to use standardized and proven technologies when storing data. A list of preferred formats can be viewed as the result of an implicit negotiation process between these stakeholders.

3.3.1. Usability versus preservation

A list of preferred formats is fruitful if the formats are usable for researchers and tractable for preservation purposes. Every archive has its community to serve, the designated community, and depending on its mission with respect to its designated community it specifies which format is preferred and which is not.

Usually, the preference is not a matter of yes or no. Formats can be preferred, acceptable, convertible, deprecated, penalised or forbidden. A good list of preferred formats is accompanied by guidelines how to convert one’s data to such a format, what to avoid and what to take extra care of, and how to check whether the result complies with the requirements.

It is not implied that data will be stored in the preferred format. It might be the case that a format is preferred because of its usability despite its weaknesses for preservation purposes. In such cases archives might migrate data in a preferred format outright into a preservation format. An example form the practice of DANS is the migration of MS Word documents to PDF.

3.3.2. Other terminology

The term *archival format* is often encountered in documentations of applications that handle data. Usually this is one of the options when saving data, and it may indicate optimisation for storage space or for short term portability or both. It often does not indicate long term usability, as this is usually not explicitly part of the policy of commercial software vendors. The other option is often the native format, which is optimised for speed. These archival formats have no connection with the preferences of an archive. Quite often the archive is less interested in economy of storage than in interoperability. Moreover, such archival formats are still vendor formats, and archives that take preservation seriously prefer open formats when available. But in the case that an archive does prefer a vendor format, it is quite possible that it will prefer its archival format as the lesser evil. This can however not be recommended as a best practice! Developing a better preservation format may be beyond the reach of the archive.

3.4. Landscape of preferred formats

Although expressing format preferences is a matter of an archive and its designated community, there is much coherence in the different positions of archives, and hence their choices resemble each other a lot. In order to identify best practices, it is a good thing to see which archive prefers what formats. Before we map out that territory, it is useful to chart the space of data kinds and their associated formats.

3.4.1. Format registries

There are sources of information on file formats: the *format registry*. A format registry collects details of file formats, often with a view to feed automatic file format recognisers, which are a very important element in maintaining a digital archive. The purpose of a format registry is not to evaluate or to choose or to prefer formats. Here we mention a few registries.

3.4.1.1 PRONOM

Created by: UK National Archives.

URL: http://www.nationalarchives.gov.uk/PRONOM

Mission: *PRONOM is being made available as an information resource for anyone who needs authoritative information about data file formats and their supporting software products, including their support lifecycles and technical requirements.*

PRONOM offers DROID as file format identification tool.

Formats can be found by searching, not by browsing.

In perusing this registry we marked some ‘shortcomings’ like: There is no entry for SPSS. There is very little information on CSV. There is virtually no info on dBase, they are mentioned, but documentation is missing. The National Archives is working hard to extend this registry, and volunteers are very welcome to contribute. We think this registry has a very good potential.

There are references to other registries, such as GDFR, UDFR, JHOVE, TOM.

3.4.1.2 GDFR (Global Digital Format Registry)

Created by: Harvard University.

URL: http://www.gdfr.info/

Mission: *The Global Digital Format Registry (GDFR) will provide sustainable distributed services to store, discover, and deliver representation information about digital formats.*

3.4.1.3 UDFR (Unified Digital Format Registry)

To be created as the merger between PRONOM and GDFR by 2010.

3.4.1.4 TOM

Created by: University of Pennsylvania

URL: http://tom.library.upenn.edu/ (broken link, no other link known)

3.4.1.5 JHOVE

Created by: JSTOR and Harvard University

URL: http://hul.harvard.edu/jhove/index.html

Mission: *JHOVE provides functions to perform format-specific identification, validation, and characterization of digital objects.*

JHOVE is a tool rather than a registry.

The mission of JHOVE is much more ambitious than that of PRONOM: format validation aims to establish the degree by which a file conforms to that format. Characterisation answers questions like: given that this file is of format F, what salient properties does it exhibit?

The flip side is that JHOVE is implemented for a limited set of formats, although it is possible to write third party plug-ins.

3.4.1.6 TrID

Created by Marco Pontello.

URL: http://mark0.net/soft-trid-e.html

Also an interesting and growing collection of file format recognition patterns. It is however unclear how stable TrID is, and what the quality is of the profiles on which the format recognition is based.

Looking at (a) to (f) the conclusion is a bit disappointing: there are at the moment no usable file format registries that have the completeness you need for preservation purposes. When facing the task to interpret an unknown format, a file format registry might provide a pointer, but most likely the critical information is to be found elsewhere.

3.4.2. Which archive prefers which formats

A list of preferred formats by an archive might be quite lengthy, with lots of additional information or guidelines, or it might be practically non-existent. We have looked at the public lists of data archives that are members of CESSDA, and the observation is that most archives do not explicitly mention preferred formats. Three archives mention a number of formats in a few sentences, such as SPSS, SAS, STATA, Access, Excel, Word, and PDF. Only one archive, the UKDA, currently has a list of preferred formats (http://www.data-archive.ac.uk/sharing/acceptable.asp), which distinguishes between preferred formats and other acceptable formats. These preferred formats are preferred for ingest because they have the potential to be more easily reduced to a software independent format, according to UKDA. This is in line with the results of the CESSDA survey in 2008 (see chapter 2). DANS will publish its list of preferred formats shortly.

These social science data archives only consider data created by statistical packages or textual data. Once other kinds of data come into play, it is advantageous to collect the experience of other users than social scientists with the formats in associated with these data kinds.

Outside CESSDA there are more developed examples of lists of preferred formats. See for example the AHDS (http://www.ahds.ac.uk/depositing/deposit-formats.htm[4]), which distinguishes between preferred formats, acceptable formats, problematic formats and even problematic aspects of formats. It organises the list in resource type. These formats were

[4] The AHDS does not exist any more, but this website is still online.

being used for ingest, not for preservation. For the type statistical dataset it prefers SPSS portable, also delimited text files with data dictionary and codebook; it accepts STATA and SAS; and it deems problematic fixed width text files without appropriate documentation. See also the detailed instructions of the Library of Congress (http://www.digitalpreservation.gov/formats/index.shtml).

UKDA (http://www.esds.ac.uk/news/publications/UKDA_DSS_QuantitativeDataProcessingProcedures.pdf) and ICPSR (http://www.icpsr.umich.edu/ICPSR/access/dataprep.pdf) issue detailed instructions for submitting data to their archives.

3.5. Wider context of digital preservation

We conclude this chapter with a selection of current work in digital preservation, especially where there are connections with file formats and metadata. We have already seen initiatives coming from individual institutions such as universities and data archives. Another category interesting: European projects in the 6th and 7th framework. There are projects in digital preservation proper, and there are infrastructure projects aimed at particular scholarly disciplines. And last but not least, there are initiatives coming from the social sciences itself to preserve and increase the usage potential of research data.

3.5.1. Digital Preservation Proper

3.5.1.1 DPE (Digital Preservation Europe)

URL: http://www.digitalpreservationeurope.eu/

DPE addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials. DPE's success will help to secure a shared knowledge base of the processes, synergy of activity, systems and techniques needed for the long-term management of digital material.

3.5.1.2 INTERPARES

URL: http://www.interpares.org/

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) aims at developing the knowledge essential to the long-term preservation of authentic records created and/or maintained in digital form and providing the basis for standards, policies, strategies and plans of action capable of ensuring the longevity of such material and the ability of its users to trust its authenticity.

3.5.1.3 CASPAR

URL: http://www.casparpreserves.eu/

How can digital data still be used and understood in the future when systems, software, and everyday knowledge continues to change? This is the CASPAR challenge. Directed at Cultural data, contemporary arts and scientific data.

3.5.1.4 PLANETS

URL: http://www.planets-project.eu/

The primary goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets.

3.5.1.5 SHAMAN

URL: http://shaman-ip.eu/shaman/node/44

Under a mid term vision, SHAMAN will design and progressive implement large-scale European-wide collections with innovative access services that support communities of practice in the creation, interpretation and use of cultural and scientific content, including multi-format and multi-source digital objects. They will be combined with robust and scalable environments which include semantic-based search capabilities and essential digital preservation features.

For the longer term, SHAMAN will develop radically new approaches to Digital Preservation, such as those inspired by human capacity to deal with information and knowledge, providing a sound basis and instruments for unleashing the potential of advanced ICT to automatically act on high volumes and dynamic and volatile digital content, guaranteeing its preservation, keeping track of its evolving semantics and usage context and safeguarding its integrity, authenticity and long term accessibility over time.

3.5.1.6 KEEP

URL: http://www.keep-project.eu/ezpub2/index.php

KEEP (Keeping Emulation Environments Portable) will develop an Emulation Access Platform to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames etc.

3.5.2. Infrastructures

3.5.2.1 CLARIN

URL: http://www.clarin.eu/

The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable. CLARIN will offer scholars the tools to allow computer-aided language processing, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences.

3.5.2.2 DARIAH

URL: http://www.dariah.eu/

DARIAH’s mission is to facilitate long-term access to, and use of all European arts and humanities data for the purposes of research. DARIAH is the digital research infrastructure that will connect scholarly data archives and repositories with cultural heritage for the arts and humanities across Europe, making scattered resources accessible through one click. DARIAH aims to create one European data area in which scholars and students can easily survey the available information in their field – data which is dependable in terms of both quality and durability. Research which builds on this data will expand knowledge and understanding of our heritage, histories, languages and cultures.

3.5.2.3 CESSDA

Url: http://www.cessda.org/ .

CESSDA is an umbrella organisation for social science data archives across Europe. Since the 1970s the members have worked together to improve access to data for researchers and students. CESSDA research and development projects and Expert Seminars enhance exchange of data and technologies among data organisations.

3.5.3. Discipline efforts

3.5.3.1 DDI (Data documentation Initiative)

URL: http://www.ddialliance.org/

3.5.3.2 DEXT (Data Exchange Tools)

Url: http://svn.opendatafoundation.org/ddidext/

3.5.3.3 MIXED(Migration to Intermediate Xml for Electronic Data)

URL: http://mixed.dans.knaw.nl/node/114

4. Recommendations for a CESSDA-wide file format registry

4.1. Introduction

In this chapter we are formulating a number of recommendations for the CESSDA RI. The recommendations deal with the following topics:

- The establishment of a CESSDA Format Registry;
- The long term preservation of data in SPSS, SAS and Stata files;
- Improve the management of upcoming diverging file formats by maintaining a preferred formats list;
- Contribute to global format registries;
- Tools for recognition of file formats;
- Tools to convert non-preferred format files into preferred formats;
- Anticipating future functions in managing file formats (technology watch).

4.2. Establishing a CESSDA Format Registry

More and more differing file formats are used in registering quantitative and qualitative data. For each old or new file format the risk of becoming obsolete must be managed and precautionary measures must be taken. For a proper assessment of a file format sometimes a lot of technical and marketing knowledge is needed. Contacts with software suppliers and standardizing committee's must be maintained. It would not be efficient when all the CESSDA organisations would perform all these tasks individually. It would be better to have some structure and cooperation in this field between the CESSDA organisations.

Our recommendation is to establish, as far as CESSDA interests are concerned, a central CESSDA format registry, supervised by a standing committee or permanent working group and maintained by contributions from individual experts from the CESSDA organisations. Since CESSDA is not alone in the file format issues, there can be many links/contributions to file format related activities outside of CESSDA. This committee could be part of the working group on CESSDA-ERIC guidelines, as proposed in the report for task 6.4 (Štebe and Dusa 2009).

This registry could function as a source of information for the CESSDA organisations in their dealings with the various file formats. Experiences could be shared about software to maintain or convert new file formats. Based on the common interests of the CESSDA organisations influence could be exerted on developments surrounding the various file formats by participating in user groups or responding to requests for comments, etc.

The standing committee for the 'CESSDA format registry could be an appointed group of five representatives from the CESSDA organisations, which takes responsibility for recording of shared knowledge within CESSDA about file formats and for formulating proposals for official standpoints of CESSDA regarding the use certain file formats. This committee coordinates the contributions of different CESSDA organisations who take responsibility to build expertise concerning a specific file format and to formulate preservation strategies and actions for that specific file format.

The tasks of the committee are

- Maintain a list of file formats that are relevant for CESSDA;
- Register per file format any collected information and possibly useful tools, and links to information elsewhere, that may be relevant to CESSDA organisations;
- Collect feedback from CESSDA organisations about file formats and tools;
- Moderate discussions within CESSDA regarding the use of certain file formats;
- Guide and acknowledge the work of the individual CESSDA organisations in building up expertise about particular file formats and contributing to the format registry;
- Represent CESSDA interests with software suppliers;
- Represent CESSDA interests regarding file formats to outside bodies, like the newly established UDFR format registry;
- Contribute to UDFR where applicable (or stimulate that activities within CESSDA contribute to UDFR);
- Advise CESSDA organisations about needed conversions and the tools to do this;
- Guide the development of an information system.

The focus of the committee is an advisory task within CESSDA and a representation of CESSDA interests to outside parties. The actual work of building up expertise concerning a particular file format and concerning conversion tools and other guidelines about the use of certain file types could be carried out by specialists within the CESSDA organisations. The results should be available to all CESSDA organisations.

Initially one of the CESSDA organisations could set up an information system on file formats. In the beginning this may be a simple spreadsheet or a text document, but an information system is needed which would make available the CESSDA preservation status of file formats to the electronic archives of the CESSDA organisations. This should be worked out further in the CESSDA RI.

4.3. Long term durability of SPSS, SAS and Stata files

For many years already the CESSDA organisations have been using files in the formats of SPSS, SAS and Stata. Because of the strategic importance of these file formats for the CESSDA community we have looked into the long term durability of these formats closer. For this we have been in contact with the support departments of the software vendors for these formats. In addition to that we have consulted Circlesys, the producer of Stat/Transfer and the digital preservation officer of ICPSR and one of the big SAS users in The Netherlands. In these contact the subject was preserving the data content of the files. We have not looked (yet) at preserving any related material, like codes, variable groups etc.

Here are some facts that emerge from these contacts:

1. SPSS and SAS have a good track record for being able to open old versions of SPSS Portable and of SAS Transport, but the companies give no guarantees as to the future support of the SPSS Portable or SAS Transport file formats. These formats are not

intended for long term preservation, but only for rather immediate exchange between different current computing environments.

2. The SPSS Portable format is not open; the SAS Transport XPORT format is open.
3. New versions of Stata have up until now supported all formats of previous version, at least to read, and the file format of Stata files is open (http://www.stata.com/help.cgi?dta).
4. Preserving statistical data for periods of 20 or 50 years in the original binary files makes you dependent on the corresponding software to exist in 20 or 50 years, or valid conversions tools to exist by then. This dependability is a serious risk, which may be a cause for data to become unusable.
5. Necessary conversions of binary files in the very far future can by no means be guaranteed in the present moment. There is always the danger of a mismatch between characteristics of different file formats.

Our conclusion from these facts is that the only sure means of preservation for the long term is converting the binary files to plain text (CSV in ASCII or Unicode). Only plain text gives the digital archive full control over the data, without being dependent on external parties.

Conversion to ASCII is also a suggestion of the SAS Support Desk (in The Netherlands), as well as a very clear outspoken recommendation by Stat/Transfer. ICPSR is following this strategy and is distributing data in plain text with added scripts (setup files) for creating binary files for the statistical software packages to work with. We are recommending conversion to plain text to the CESSDA organisations. See http://www.icpsr.umich.edu/ICPSR/help/datausers/usingdata.html and http://www.icpsr.umich.edu/DP/ . ICPSR has developed a fully automated conversion process for SPSS files.

The best moment to export the data to plain text is as soon as possible after having received the files, because only then you are able to confer with the data producer or with the software supplier in case of any problems. All the mentioned statistical packages here support export to plain text and import from plain text, or there are easy available utilities to do this, like Stat/Transfer.

There are some remaining points of attention in dealing with conversion to plain text:

- The plain text version preferably should be in Unicode encoding to avoid any misunderstanding of (for instance) extended ASCII characters and the like. According to our information SAS supports exports to and imports from Unicode, and Stat/Transfer will support UTF-8 in the next release. The issue of codepages is sometimes complicated and we have not found time to really sort this out to our own satisfaction. We recommend to look into this further in the coming years.
- There should be more independent quality controls over plain text exports or recreated binary files than we see now in practice. Possibly the Universal Numerical Fingerprints

(UNF) (http://thedata.org/citation/tech and http://cran.osmirror.nl/web/packages/UNF/index.html) can render services here, or some simple generally usable scripts can be developed.

- More and more data is stored in XML files. The advantage over plain text like CSV is that it is possible to give more meaning to the data, and it is still possible to read and fairly easy to manipulate the data without being dependent on specialized software. For instance, a combination of data and DDI documentation in one file is possible with XML. The MIXED project run by DANS is working in this direction (as was the DeXt project). See http://mixed.dans.knaw.nl/files/file/white_paper2.pdf for the Mixed project initiation. The MIXED website http://mixed.dans.knaw.nl is being renovated.

4.4. Proposal of a preferred formats list

In Appendix A to this report we present a current list of preferred formats as compiled by DANS. This list is intended for use in both the social sciences and the humanities and should be considered as a starting point for discussion as DANS is an archive for research data in the Social Sciences and Humanities: the latter in particular history and archaeology. From the perspective of file formats this means that DANS has a long experience with statistical data (the Steinmetz Archive) and historical data like databases, spreadsheets and to a lesser degree texts (Netherlands Historical Data Archive). Especially in the field of archaeology there have been many additions lately (images, 3D compositions, geographical information and databases). The basic philosophy has always been to put much effort in saving original files also in formats that were considered at the time more durable than the original files, like in plain text and/or in PDF for word processor document, or in CSV for databases, etc. The durable file versions offer more certainty of the data/content being at least readable in de far future, but reusing the data/content with all the functionality of the original file formats is often hampered. Only in recent years we see some durable formats appearing with rather rich functionality, for instance the Open Document format.

A list of preferred formats is by no means a static thing. Although the landscape of emerging and disappearing file formats is not changing quickly, it is changing. We advise a yearly evaluation of all file formats that are relevant to an archive. Major policy changes regarding the status of certain file formats with regard to long term durability happen every 5 to 10 years or take even longer. The file formats common in statistical analysis have remained remarkably steady over the past decades. In contrast word processor documents in the DOS and Windows PC environment have developed from Wordstar and the likes to Word Perfect, to MS Word in various versions, to possibly Open Document Format in the future, with many less dominant formats lingering along. There is a good chance that someone who started personal computing in the first half of the nineties has files on his computer that he can’t read with the software of today. The preferred formats list in Appendix A also starts with a short motivation for focussing on preferred formats.

Having a preferred formats list would make clear which file formats will be actively supported by the archive. It gives direction to the data producers. Many other file formats however exist, so there is need for tools to convert non-preferred formats to preferred formats and instructions on how to do this in the best possible way. Information is needed about potential conversion problems. See the next section where these tools and instructions are being described for a start.

4.5. Tools to convert non-preferred format files into preferred formats

In addition to the proposed Preferred Formats list, we have compiled a list of tools to convert non-preferred formats to preferred formats, see Appendix B. This list is also no more than a starting point for discussion within CESSDA. Is such a list needed in the CESSDA RI? If so, how shall it be maintained? What different categories should be discerned? We suggest that this will be coordinated by the body that maintains the CESSDA Format Registry (see section 4.2).

4.6. CSV export problems for databases and directions for solutions

At present, it appears to be difficult to decide on a satisfactory solution for preserving databases in a durable and accessible format. The DANS Preferred Formats list recommends an export of the individual tables in a database to Comma Separated Values (CSV).

The CSV format is clearly structured, easily readable, in widespread use and well-supported by software. This makes it an attractive candidate for preservation, however, there is no general standard specification for CSV and differences between CSV formats are easily found. Software, system settings and regional settings define the CSV export of a database table or spreadsheet. Export guidelines and definitions of the exact desired CSV encoding are required to ensure the storage of all data in a uniform manner so that the CSV export can be regarded as a good preservation format. To avoid problems with diacritics, punctuation etc. an export to Unicode is advised.

A suggestion to maintain a uniform standard is to write a detailed export function/script to use for all exports. Ideally, such an export script should cover all expected data types (text, memo, integer, double integer …). The following issues need specific attention when writing the export code:

Number types: Depending on the computer's regional settings, number fields in a table may be displayed and exported with commas instead of dots as dividers. A CSV export will enclose fields with commas in double quotes. It is regarded as a CSV standard to not have numbers in double quotes.

We suggest exporting numbers as: unquoted digits, using decimal dots in case of dividers.

Decimal digits: Microsoft products, including access, may automatically round up numbers with decimal dividers to the second digit after the divider, depending on the data type. This can result in loss of essential data. For example, if a coordinate is written as x= 123.456, the value will be rounded up to 123.46.

The field size can be set to 'decimal' and the scale to the required amount of decimals after the divider to keep the original value.

The problem is that a Microsoft product may not export a data type as the data type it should be. The Access export wizard will regard all number types as a decimal with two digits after the divider. An export script within a Microsoft product (i.e. Access) can bypass this problem by specifically stating that each data type should be exported as such: Decimals should be regarded as Decimals.

Dates: There are different variations of denoting dates, in the order of days, months, years, and in the use two or four digits for years. How to treat dates? Translate all dates to the same notation? Regard dates as text, between double quotes?

We suggest to export all dates as: yyyy-mm-hhThh:mm:ss (unquoted). See also ISO 8601.

Booleans: How to export columns where the field can be either 'Yes' or 'No', tagged or untagged? Solutions could be text values of "Y" or "N", or number values of 1 or 0.

We suggest exporting Booleans as: 1 or 0 (unquoted).

Currency: Currency symbols will be modified to (local) regional settings, which could turn dollars to Euros, etcetera. We suggest exporting currency as: value, without currency symbol. The data type should be changed back from value to currency whenever the table will be imported, so this modification should be clearly registered in a metadata document.

Microsoft Access can be used to create an export form in an empty database; single databases can then be imported in the database and the form can be copied into other Access databases.

If the exported tables will be imported again, the data should be replicated with the exact same precision as it was contained in the original database or tables. To ensure that the data can be reproduced as such, the data types, table structures and any codepages need to be documented in metadata files.

The MIXED project (Migration to Intermediate XML for Electronic Data) will eventually allow for proper import and export of databases: see http://mixed.dans.knaw.nl/

4.7. Future preservation management functions within an archive

In this section we mention a number of miscellaneous preservation issues that may require the attention of the CESSDA organisations at this moment or somewhere in the future. Possibly the CESSDA organisations already have dealt with these issues in their operations, or are planning to deal with them. Cooperation within CESSDA seems often possible.

4.7.1. Quality guidelines & Preservation Policy documents

To assist the management of digital archives several quality guidelines have been proposed:

- Trustworthy Repositories Audit & Certification (TRAC) http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91 . For current review process see http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/WebHome . On ISO standardisation of TRAC(see http://www.digitalrepositoryauditandcertification.org/.

- NESTOR website http://www.langzeitarchivierung.de , criteria version 1 http://www.nbn-resolving.de/?urn:nbn:de:0008-2006060703 and version 2 in German http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_08.pdf .
- Data Seal of Approval, website http://www.datasealofapproval.org/ , criteria http://www.datasealofapproval.org/sites/default/files/DSA-Assessment%20form%20v2-2.doc
- Reference Model for an Open Archival Information System (OAIS), http://public.ccsds.org/publications/archive/650x0b1.pdf

By setting up or maintaining a preservation policy document the staff of a digital archive can evaluate their current position against these quality guidelines and lay out a plan for future developments.

Here are some examples of preservation policy documents: http://www.data-archive.ac.uk/news/publications/preservationpolicy.pdf and http://www.icpsr.umich.edu/DP/policies/dpp-framework.html .

4.7.2. Establish quality guidelines / scientific criteria for creation of new data types

The CESSDA organisations have a long history of keeping and distributing statistical files. For the creation of these files best practices have evolved over time, which more or less guarantee the quality of the data in these files. See for instance the Guide to Social Science Data Preparation and Archiving, 3rd Edition by ICPSR, http://www.icpsr.umich.edu/ICPSR/access/dataprep.pdf . There is a body of knowledge about setting up and conducting surveys in a scientifically reliable way. This knowledge is strongly related to the methodology of the social sciences, in particular regarding data collection.

It would be a good thing if similar quality safeguards would be available for other types of data that is distributed by a digital archive, especially in the humanities where there are less explicit and advanced methodologies. Other types of data are for instance:

- Interviews;
- Formulas in spreadsheets;
- Ethnographic descriptions;
- Images;
- Geographic information.

No established quality criteria for the new types of files exist. This could give an uneasy feeling to long standing respected archives. We have not found time at the moment to research this further, but we would like to give CESSDA into consideration to start collecting best practices for the creation of these new types of data. For instance, for properly conducting interviews there are recommendations in the area of Oral History (http://www.h-net.org/~oralhist/).

4.7.3. Version management and provenance data

In preserving data different versions of the same original files will be created, like exports to plain text of migrations to new file formats. Over a long period of time even a chain of file versions can emerge that are all linked to one original deposited file. The supporting archive information system should be able to record the relations between these files and the accompanying metadata (see also the OAIS reference model). For each newly created version information must be recorded about the versions of the tool(s) with which the file version is created and with which values of parameters if any the conversion is performed, and the precise file format that is created. This 'provenance data' is necessary to be accountable for all the manipulations on deposited files. This information is also necessary if anytime later problems may be detected in the uses tools and decisions have to be taken about repeating or fixing certain migrations. For each used (sub)version of a tool information must be maintained about dependencies to other system software and any known limitations and defects.

4.7.4. Work towards setting up a migration framework

The Planets http://www.planets-project.eu/ project is working hard to develop a strategy model and generic tools to keep the files in a digital archive in usable up-to-date file formats. This year they are rolling out Testbed, a tool to make it easy for archivists to test conversion services that are becoming available on the internet (http://gforge.planets-project.eu/gf/project/ptb). Another Planet tool is Plato a decision support tool that implements a solid preservation planning process and integrates services for content characterisation, preservation action and automatic object comparison in a service-oriented architecture to provide maximum support for preservation planning endeavours (http://www.ifs.tuwien.ac.at/dp/plato/intro.html).

It remains to be seen whether or to what extend the deliverables of the Planets project will be usable in the technical environments of the CESSDA organisations.

In our opinion a fully developed migration framework to manage file formats will contain at least these functions:

1. On ingest
    a. Determine detailed file formats of files (technical metadata).
    b. Determine whether the files fully comply with the detected format.
2. Monitoring
    a. Maintain registry of file formats and of possible preservation actions, and link this information to preservation policy of the archive
    b. Periodically assess the need to take preservation action, or to redo some already performed migrations
    c. Process new/adjusted information about file formats and migration tools
3. Migration
    a. Perform preservation action (migration) + record provenance information
    b. Assess quality of the performed migrations (compare essential properties)
    c. Present appropriate file version to users of the digital archive
    d. Maintain information about available migration services on the internet or with partners/service providers.
    e. Maintain an in house set of migration tools

Such a framework cannot be developed overnight. It may take some years to develop it. CESSDA organisations can cooperate in developing parts of the framework, or in sharing experience with using components that have been developed elsewhere (for instance in the Planets project).

4.7.5. Enhanced publications

A new type of publication is emerging, called ‘enhanced publications’, see http://www.driver-repository.eu/Enhanced-Publications.html. With this term publications are meant that combine digital artefact from various sources. For instance a word processor document may link as embedded objects images / spreadsheets / audio / video files from the same archive or from other archives. This means the referenced objects must have stable (persistent) references and the access path to the involved digital archives must allow for these objects to be read in the context of the enhanced publication.

The rationale behind “enhanced publications” is threefold: link between publication and (1) research data in order to give reader access to research data on which research is based (2) extra material to illustrate or elaborate on the research (3) “post-publication” data: comments and assessments (see: “Report on enhanced publications state-of-the-art” Deliverable D4.1 of the DRIVER project, see http://www.driver-repository.eu.

5. Summary of Recommendations

In this report the following recommendations are made:

1. A clear set of guidelines should be created, the "CESSDA-ERIC requirements for operating trusted data repositories", implying a CESSDA-ERIC 'seal of approval'. This should include use of the OAIS reference model.
2. A reduced version of these guidelines, adapted for small-scale institutes, should be made.
3. The creation of assessment procedures for CESSDA RI partners, appropriate to the level of service which the partner is going to provide.
4. These guidelines should include the relevant legal framework needed for digital preservation; including both national legislation concerning the protection of personal data, intellectual property rights (legality of copying data files for preservation purposes) and "codes of conduct" for the exchange of knowledge and information.
5. Orientation towards the newest developments in the Creative Commons Movement is strongly advisable, in particular on data.
6. Following the developments in the Open Access Movement is advisable.
7. The establishment of a central CESSDA format registry, supervised by a standing committee or permanent working group and maintained by contributions from individual experts from the CESSDA organisations. This committee could be part of the working group on CESSDA-ERIC guidelines, as proposed in the report for task 6.4 (Štebe and Dusa).
8. The central CESSDA format registry could be linked to or form part of global file format registries.
9. The set-up of a CESSDA RI information system on file formats as an initial step towards a file format registry.
10. The only sure means of preservation for the long term of binary files is converting them into plain text, preferably in Unicode (or ASCII, CSV).
11. More independent quality controls over plain text exports are advised.

For the full lists of preferred format lists, conversion tools and file format identification tools we refer to the appendices of this report.

6. References

Beagrie Neil, Julia Chruszcz and Brian Lavoie (2008): Keeping Research Data Safe. JISC Report http://www.jisc.ac.uk/publications/documents/keepingresearchdatasafe.aspx

Beedham Hilary et al. (2005): *Assessment of UKDA and TNA compliance with OAIS and METS standards* (UK Data Archive 2005) http://www.data-archive.ac.uk/news/publications/oaismets.pdf

Blank, G. & Rasmussen, K.B. (2004): The Data Documentation Initiative: The Value and Significance of a Worldwide Standard. *Social Science Computer Review*, Vol. 22, No. 3, Fall 2004 307-318

CESSDA Digital Preservation report (2009): report of the CESSDA subgroup at the Digital Preservation Workshop - January 30th 2009 - The Hague. http://www.datasealofapproval.org/?q=node/3

Chronopolis: Federated Digital Preservation Across Space and Time project http://chronopolis.sdsc.edu/

Corti Louise (2007): 'Re-using archived qualitative data – where, how, why?', *Archival Science*, Vol. 7, Number 1 / March, 2007, pp. 37-54, DOI 10.1007/s10502-007-9054-6

Creative Commons: http://creativecommons.org/

Doorn Peter (2004): 'Research Data Archives and Public Electronic Record-Offices: What can we learn from each other?' in: Peter Doorn, Irina Garskova and Heiko Tjalsma (eds.), *Archives in Cyberspace. Electronic Records in East and West* (Moscow 2004) 98.

Doorn Peter and Heiko Tjalsma (2007): 'Introduction: archiving research data', *Archival Science*, Vol. 7, Number 1 / March, 2007, pp. 1-20, DOI 10.1007/s10502-007-9054-6

Data Seal of Approval: http://www.datasealofapproval.org/

Fábián Zoltán (2009): Self-assessment procedures for the CESSDA RI infrastructures. Interim report from Task 6.2 WP6, 0.9 draft

FORS: http://www.unil.ch/fors/page60005_en.html

FSD: http://www.fsd.uta.fi/tietoarkistolehti/english/22/paakirjoitus.html

JISC Standards Catalogue, http://standards-catalogue.ukoln.ac.uk/index/OAIS

LOCKSS: http://www.lockss.org/lockss/Publications

OAIS Blue Book (2002): http://public.ccsds.org/publications/archive/650x0b1.pdf

Ockerbloom John Mark (2008): What repositories do: The OAIS model, http://everybodyslibraries.com/2008/10/13/what-repositories-do-the-oais-model/

Open Access: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

O'Neill Adams Margaret (2007): 'Analyzing Archives and Finding Facts: Use and Users of Digital Data Records'. *Archival Science*, Vol. 7, Number 1 / March, 2007, pp. 21-36, DOI 10.1007/s10502-007-9054-6

Preserving Digital Information (1996): Report of the Task Force on Archiving of Digital Information
Commissioned by The Commission on Preservation and Access and The Research Libraries Group May 1, 1996 http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

Qualidata: http://www.esds.ac.uk/qualidata/

Rasmussen Karsten Boye and Grant Blank (2007): 'The Data Documentation Initiative: A Preservation Standard for Research', *Archival Science*, Vol. 7, Number 1 / March, 2007, pp. 55-72, DOI 10.1007/s10502-007-9054-6

Ruusalepp Raivo (2009): Building Blocks of a Certification Process: DRAMBORA and TRAC. Presentation given at Digital Preservation Workshop - January 30th 2009 - The Hague. http://www.datasealofapproval.org/?q=node/3

Schumann Natasha (2009): nestor. Short reflections on trusted digital repositories and building blocks for a certification process. Presentation given at Digital Preservation Workshop - January 30th 2009 - The Hague. http://www.datasealofapproval.org/?q=node/3

SPSS: http://www.SPSS.com/corpinfo/history.htm

Štebe Janez and Dusa Adrian (2009): Task 6.4 Report: Recommendations concerning best practices WP6, Draft v04

Vardigan Mary and Cole Whiteman (2007): ICPSR meets OAIS: applying the OAIS reference model to the social science archive context pp. 73-88, DOI 10.1007/s10502-007-9054-6

Woollard Matthew (2009): Presentation given at Digital Preservation Workshop - January 30th 2009 - The Hague. http://www.datasealofapproval.org/?q=node/3

NB. References in the chapters 3 and 4 and the appendices have mainly been kept in the body of the text, to enable efficient reading.

7. Glossary

AHDS – Arts and Humanities Data Service
CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval project
CESSDA – Council of European Social Science Data Archives
CLARIN – Common Language Resources and Technology Infrastructure
DARIAH – Digital Research Infrastructure for the Arts and Humanities
DARM – Data Activities Reference Model
DDI - Data Documentation Initiative
DDQ - Digital Document Quarterly
DExT - Data Exchange Tools and Conversion Utilities
DPE - Digital Preservation Europe
DRAMBORA - Digital Repository Audit Method Based on Risk Assessment
DSA – Data Seal of Approval
ERI – European Research Infrastructure
ERIC – European Research Infrastructure Consortium
GDFR - Global Digital Format Registry
ISAD(G) - General International Standard Archival Description
ISO - International Organization for Standardization
KEEP - Keeping Emulation Environments Portable project
LOCKSS - Lots of Copies Keep Stuff Safe project
MIXED - Migration to Intermediate Xml for Electronic Data project
NESTOR - NEtwork of Expertise in Long-term STOrage of Digital Resources
OAIS - Open Archival Information System
OAI-PMH - The Open Archives Initiative Protocol for Metadata Harvesting
SDS - Standard Study Description Scheme
TRAC - Trustworthy Repositories Audit & Certification: Criteria and Checklist
UDFR - Unified Digital Format Registry

Appendix A. - Preferred formats list of DANS

A.1. Introduction

All formats of digital files stand the risk of becoming obsolete in the future. Obsolete sometime in the future means that the then current software is not able to represent and use the content of the file in the way as it was meant at the time of creation. This is a big risk that needs to be taken very seriously. When a file format becomes obsolete there are two possibilities: the first is that there is nothing we can do to remedy the situation. The file is unusable and that's it, lost forever. The second possibility is that there is software available which can convert the obsolete file in a trustworthy manner into a then current file format. If conversion is possible, then the quality of this conversion becomes an important issue. Are all the essential properties of the original file maintained in the converted file (and can this be verified in an automated manner)? For the files in the archive of DANS it is usually more important to maintain the data content properly, than to maintain the way the data is presented.

DANS maintains a digital archive, called EASY, for research data in the Humanities and Social Sciences. This document is limited to file types that we encounter while maintaining this digital archive. Being a small organization we very much feel the need to restrict ourselves to just a few strategic file formats to concentrate our preservation efforts on. This document mentions the file formats in which we have at this moment the highest confidence regarding long term usability. Without giving any guarantees we can state that we will do our utmost best to preserve the content of files in these formats, in the case that any of these formats might become obsolete. While saying this, we realize very well that in this we will be very dependent on the availability of tools produced by commercial companies or by government sponsored international projects.

It is impossible to predict for a given file format whether or not it will be current in 10 years, or in 50 or 100 years. In general the file formats that are used much and of which the specification are openly published and free of rights, have a higher chance of surviving. Many other aspects can be considered in judging the chances of a certain format to survive, see for instance the format evaluation form developed by the KB[5]. To compile this restricted list we have done research on the internet and we have taken into account our own experience with the file formats of research data up until now.

We urgently ask the depositors the use the formats we recommend here to submit data in our archive. If data is originally created in another format, then we also welcome the original format because probably in the near future some people can benefit from this original format too. We think the creator of the data is himself the best person to judge a conversion from an original format to a preferred format.

[5] See http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf

There is a tradeoff between ‘the highest chances of preserving at least some of the data content of a file’ and ‘the ability to reuse the data with all the functionality of the original file’. For instance, converting a word processor document to plain text augments the chances of preserving at least some of the content in the distant future, but reduces strongly the functionality of headers, footnotes, tables, etc. When you think of this you come to the conclusion that it can be very sensible to actually store one original file in several different formats. In the list below we indicate this with each file type.

At DANS we make distinction between preferred formats and convertible formats.

While maintaining a digital library with files in any of the formats mentioned in this document, it is important to periodically check the prospects of each format. When a preferred format, in spite of all the good hopes, is in danger of becoming obsolete, conversions must be planned of the endangered preferred format files and/or of the corresponding convertible format files to a new chosen preferred format. When a convertible format is in danger of becoming obsolete, this might be a trigger to evaluate one more time the chosen preferred format and the quality of the performed conversions. When new releases become available of used conversion software, this might be a reason to consider redoing conversions. A constant alertness is necessary.

PDF/A is prominent in the proposed preferred formats list. PDF has been around for a long time. It is available on many platforms. With PDF/A the format is open. We have a high trust in the long term durability of PDF/A. We cannot imagine ‘the world’ letting go of PDF/A without having a sound and solid migration path to a sound and solid alternative. In fact, for many types of content, PDF is the replacement of good old paper prints. Where possible we advise to add document structure tags to improve the re-use of parts of the text in a PDF/A file.

A.2. Format categories

A.2.1. Preferred Formats

Preferred formats are formats that have, at this moment and to our best knowledge, the best chance of ‘surviving’ in the distant future (20 or more years away) or having then good conversion software available which preserves the essential attributes of the file. Preferred formats may have less functionality than convertible formats, or other current formats. Please check the result of conversions to preferred formats.

A.2.2. Convertible Formats

With convertible formats we mean formats for which there is at this moment an acceptable conversion means to a preferred format and which we ourselves are able to perform (software available). Depositors are strongly encouraged to convert data in convertible formats to preferred formats themselves and validate the result. Both file formats can be deposited and will be preserved by DANS. DANS gives no guarantees as to being able in the future to process convertible formats.

A.2.3. Other format types

Apart from “preferred formats” and “convertible formats” there are some other types of formats that are used in discussions about archiving. We mention them here shortly to avoid confusion and unnecessary questions. With “archival formats” formats are meant that an archive makes for its own purpose for reasons of efficiency or durability and which are used to create future “distribution formats”. Distribution formats are formats an archive uses to distribute files. These may differ from formats in which files are received. “Presentation formats” are formats used to present data for instance on web pages. “Original formats” are the formats in which data is deposited into the archive. The preferred formats as mentioned in this document often fulfill many or most of these functions and are an attempt to merge these differences.

A.3. Preferred formats per file type

A.3.1. Word processor documents

A.3.1.1. Fixed text

- Preferred: PDF/A
- Convertible: ODT, DOC, DOCX, RTF
- Rationale: The PDF format has become a main text format through worldwide use. PDF is meant to store all content of a document for display independent of software or hardware. The PDF/A version is designed for long-term preservation. All fonts and images are embedded in the file itself, so that the file stores the information necessary to reproduce the document. PDF/A has been globally adopted and seems to have firmly established itself as an archiving format for text. In the case of fixed text no processing later on is necessary, so we can skip the word processor formats.

A.3.1.2. Reusable text

- Preferred: PDF/A and ODT (both)
- Convertible: DOC, DOCX, RTF
- Rationale: ODT has at this moment the best mix of durability and reusability, but the Open Document format is rather new (from the perspective of a long term archive) ... As a matter of precaution we advise to create a PDF/A version also.

A.3.2. Plain text

- Preferred: UNICODE with Byte Order Mark (UTF-8, UTF-16 or UTF-32)
- Convertible: ASCII, Extended ASCII with documented codepage
- Rationale: we advise the Unicode variant to be sure that all the characters used have the proper (intended) meaning in all computing environments.

A.3.3. Presentation (slides for projection)

- Preferred: PDF/A and ODP (both)
- Convertible: PPT, PPTX

- Rationale: ODP has at this moment the best mix of durability and reusability, but the Open Document format is rather new(from the perspective of a long term archive) ... As a matter of precaution we advise to create a PDF/A version also.

A.3.4. Still Images

A.3.4.1. Raster images

- Preferred: JPEG, TIFF
- Convertible: all current formats
- Rationale: we think JPEG is established enough to serve as a preservation format. An advantage is the support by web browsers, a disadvantage is the compression used if an image is saved to JPEG from another format, possibly resulting in quality loss. For high quality graphics TIFF can be used as an archiving format. If TIFF is used as an archiving format, JPEG could be simultaneously adopted as a presentation format instead.

A.3.4.2. Vector images

- Preferred: PDF/A, SVG. Note: in case of conversion to SVG please check the result carefully.
- Convertible: AI, EPS
- Rationale: PDF/A and SVG are open standards. SVG has only been in development since 1999 and software support is relatively limited. Internet Explorer requires a plug-in to show SVG images. Still, the XML-based structure and the capability of SVG to define and embed fonts, metadata and links make SVG a strong archival format, recommended by the World Wide Web Consortium. SVG also does not have the issue of version incompatibilities that is apparent in other formats such as EPS.

A.3.5. Moving Images

- Preferred: MPEG-2, MPEG-4 H264, lossless AVI (Windows), QuickTime DV (MAC)
- Convertible: (no formats specified)
- Rationale: MPEG-2 is most used for distribution at the moment, but is in the process of being more and more replaced by MPEG-4. Both formats are open, but use lossy compression. Unfortunately there is no common open format for high quality video. The audio video industry is divided into Windows and Mac camps. From both side we have chosen a common high quality video format. There are so many formats for moving images, that we find it hard to pinpoint specific convertible formats.

Note: DANS has no experience with moving images at the moment, but we are preparing for the intake of several hundred interview recordings.

A.3.6. Audio

- Preferred: MP3, WAV (high quality Windows), AIFF (high quality MAC)
- Convertible: (no formats specified)
- Rationale: the rationale for audio is very similar to the rationale for moving images.

Note: DANS has no experience with audio files at the moment.

A.3.7. Spreadsheets

- Preferred: PDF/A and ODS (both)
- Convertible: XLS, XLSX
- Rationale: ODS has at this moment the best mix of durability and reusability, but the Open Document format is rather new(from the perspective of a long term archive) ... As a matter of precaution we advise to create a PDF/A version also.

A.3.8. Database

- Preferred: unfortunately there exists at the moment no generally usable preferred format, but we advise archives to create CSV files from the tables of deposited databases.
- Convertible: DBF, MDB, ACCDB
- Rationale: database formats like dBase or Microsoft Access are very widely used. Unfortunately the prospects for long term usability are not good, but there is at the moment no other candidate for the role. The export functions to CSV should be used carefully to avoid problems with the text representation of data elements and to avoid problems with international characters. That is why we advise that the export to CSV should be done by the archive. If someone needs help in making a CSV export, please ask DANS for advice. Depositors should provide documentation about the relations between the tables and about each column in a table (meaning of the values in the column, data type), and count of the rows in the tables and any other checksum possible (like totals of numeric data).

Notes: with databases we mean the, often single file, databases that can easily be copied or uploaded to another environment. We are at this moment developing more experience and tools with preserving databases, see http://mixed.dans.knaw.nl .

A.3.9. Statistical data

- Preferred: SPSS portable, SAS Transport, STATA DTA
- Convertible: (no formats specified)
- Rationale: there seems to be a very widespread agreement among statistical data producers and consumers that SPSS, SAS or Stata are the ones to use. SPSS and SAS have a very long record of serving management of digital data.

A.3.10.GIS (Geographic information system)

- Preferred: Mapinfo Mid/Mif
- Convertible: Mapinfo TAB, ESRI Shape files (shp)
- Rationale: TAB and SHP are binary files, TAB may include obsolete path references, SHP is dependent on ESRI software, making these main GIS file formats not suitable for preservation. The inter-related files MID and MIF are text exports of GIS files meant to store all information for import in various GIS software. The files are clearly structured and well supported.

Note: we expect GML to develop into a good preservation format for GIS.

A.3.11.CAD

- Preferred: DXF version R12
- Convertible: DWG
- Rationale: DXF has been designed to store CAD files in an accessible format as opposed to the AutoCAD propriety binary format DWG. DXF is widely supported by various CAD and GIS software. There are different DXF versions in existence; there is more support for the R12 version than there is for later versions. A problem with DXF is that the development of the DWG format continues to the point that DWG may grow to contain features that cannot be written to the DXF export. Advanced CAD features might be more expected in certain disciplines than in others, like in architecture. The output needs to be checked for loss of data.
- Note: we expect GML to develop into a preservation format for CAD and GIS alike. Unfortunately, the GML format is still in development at present, and can not yet be adopted as a preservation format. We suggest archiving CAD files as DXF R12 until GML can be used.

Appendix B. - Conversion Tools

The purpose of this appendix is to provide a depositor or data archive with suggestions for converting data files which are not stored as durable archival formats, into their corresponding Preferred Formats as defined by DANS (see the appendix of the Preferred Formats list). By no means is this chapter a complete guidebook for file format conversions. There are more tools available than listed below; any suggested software are merely examples.

This chapter might be seen as a starting point for an internal CESSDA discussion, as is the proposed preferred formats list. It could benefit all CESSDA members to create a collective, more complete list of tools.

DANS has first-hand experience with most file categories but not all: at the moment we have limited experience with audio/video files and we have only recently revised our own strategy of storing statistical data. Apart from that, it should be pointed out that most manual format conversions at DANS are being done under Microsoft Windows; we do not provide suggestions for conversions on other platforms for some of the file categories, although at the enterprise level we are developing conversions that run under Linux.

B.1. Word processor documents

Fixed text: ODT, DOC, DOCX, RTF => PDF/A

Reusable text: DOC, DOCX, RTF => PDF/A and ODT

All Microsoft Office 2007 products have a ‘save as PDF’ feature. Options within this feature include saving to the preferred PDF/A format and a check box for placing document structure tags for accessibility; it is recommended to tick this check box.

If Adobe Acrobat is available, the additional Acrobat PDFMaker tool can be downloaded from the Adobe website free of charge. The PDFMaker installs itself as a printer. All printing assignments can then be sent to the ‘Adobe PDF’ printer, which will print the document into a new PDF file. Additionally, the tool creates a ‘Convert to Adobe PDF’ option when right-clicking a document, or a selection of several document files, in Windows Explorer, thus allowing batch conversions.

An Adobe PDF drop-down menu will also be created in all Microsoft Office programs.

Be warned that sending a document to the Adobe PDF printer will not preserve any hyperlinks other than a full web address. The ‘Convert to Adobe PDF’ option and the features in the Adobe PDF drop-down menus *will* preserve hyperlinks and should be used instead.

After installing the PDFMaker tool for the first time, several settings need to be changed before the documents will be saved as the desired PDF/A format:

-The Printer Preferences of the Adobe PDF printer should have ‘PDF/A-1b:2005 (RGB)’, size A4, as the default setting.

-Adobe Acrobat preferences includes a ‘Convert To PDF’ category: all format settings under this category need to be set to ‘PDF/A-1b:2005 (RGB)’.

-Microsoft Office software also needs to be adjusted: the Adobe PDF drop-down menus includes the ‘Change Conversion Settings’ option, wherein the default setting must be set to ‘PDF/A-1b:2005 (RGB)’.

If all fonts in a newly created PDF file are listed as Embedded Subsets under document properties in the PDF viewer, the document has been successfully converted to PDF/A. This is not a foolproof check; documents which do not have their fonts listed as PDF/A may still be PDF/A. There are verification tools available in case of doubt, such as the LuraTech PDF/A Validator.

A document can be saved from Open Office in the ODT format. No additional tool is necessary. Open Office is downloadable from the Internet free of charge.

Service Pack 2 for Microsoft Office 2007 contains Microsoft’s first native implementation of the Open Document Format. There is presently some discussion on the issue of Microsoft’s use of formula specifications for Open Document Formats and the degree of interoperability of SP2-created Open Document files.

The Macintosh comes with a PDF printing feature similar to the one of the Adobe PDFMaker as described above; take care that this feature also shares the problem that hyperlinks will not be preserved.

B.2. Plain text

ASCII, Extended ASCII => UNICODE with Byte Order Mark (UTF-8, -16 or -32)

If a text document is opened in Windows Notepad, it can be saved with the encoding specified as UTF.

The script ASCII2UC.VBS to convert ASCII to Unicode is available at http://www.robvanderwoude.com/type.php. The usage with: “CSCRIPT.EXE //NoLogo ASCII2UC.VBS ascii_file unicode_file” can be applied to batches if written in a .BAT command file, taking care of individual filenames.

A user-friendly tool for batch conversions of text files in Microsoft Windows is Sisulizer’s Kaboom, available at http://www.sisulizer.com/kaboom/index.shtml. The free download includes a Multi-Converter with a ‘drag and drop’ interface. It recognizes the format of text files dragged into the menu and will be able to export files as UTF-8 or -16. The export should be specified to ‘Write BOM’ (Byte Order Mark). If desired, the Converter keeps the original files as .bak backup files. A command-line option is available for users who make a donation to Sisulizer.

B.3. Presentation

PPT, PPTX => PDF/A and ODP

Like word processor documents, Office 2007 can save a Powerpoint file as a PDF/A. Likewise, a Powerpoint file (or batch of files) can be printed to the Adobe PDF printer just like documents. Please refer to the section on **Word Processor Documents** above.

Documents are saved to ODP if they are saved in Open Office. This can also be done in Microsoft Office 2007 with Service Pack 2, as with ODT (see **Word Processor Documents)**.

B.4. Images; raster images

All current formats => JPEG, TIFF

The file formats JPEG and TIFF are so common that it can be expected that any raster graphics software can save the images it can open in JPEG and TIFF. Limitations may be more present in the import capabilities of the software; for example, the program Paint (included with Microsoft Windows systems) can only open six different raster image formats.

Adobe Photoshop, Corel (/Jasc) Paint Shop Pro and IrfanView are only three examples of software that can open a large variety of raster image formats. Both of these programs also allow for batch conversions. Batch conversions with Adobe software are done via the use of 'Actions'; a feature that allows the user to record a specific procedure, which can then be applied to other files or folders.

Some raster images may require an extra change before they can be saved as the desired format. An Adobe Photoshop action can include many kinds of adjustments, for example changing the image size, rotating the image, changing the resolution, changing from Greyscale to Bitmap, changing from Indexed color to RGB or CMYK, changing the from 16-bits to 8-bits, …

A custom pixel aspect ratio should be set to 'Square'.

Macintosh systems include Preview, which can be used for converting many different image file formats. The program does not include a batch feature.

The open-source software GIMP (GNU Image Manipulation Program) is available for many operating systems and is included on many Linux systems. GIMP is sometimes regarded as a substitute for Adobe Photoshop, but significant differences include batch processing options: GIMP requires basic programming knowledge to automate features.

Please take care that the chosen software will save an image to JPEG in the maximum quality.

B.5. Images; vector images

AI, EPS => PDF/A, SVG

Most vector graphics editors will be able to export a vector image to the SVG format. Adobe Illustrator can be used for batch conversions, with 'Actions' as with Photoshop for raster images (see **Images; raster images**).

Older vector images may contain specific fonts that may not be recognized by the Illustrator software. The text will be projected but not automatically exported, unless the fonts are changed to a general font: ‘Find Font’, change all (for example) to ‘Arial’.

A vector image can be printed to the Adobe PDF printer like documents, as described above (see **Word Processor Documents**). Take care that the Illustrator print area covers the entire image: if this is not the case, aspects such as the image ‘Size’ and ‘Orientation’ need to be changed. These settings need to be adjusted in both ‘Document Setup’ *and* the Illustrator Print menu.

B.6. Moving Images / Audio

(unspecified) => MPEG-2, MPEG-4 H264, AVI, DV

(unspecified) => MP3 (256 kbps), WAV, AIFF

Specialized editing software such as Final Cut Pro and AVID are used by professional companies dealing with audio and video files. These programs have conversion capabilities, but the software may be too complex and/or costly for use outside of the professional audio/video world.

Other software can be found on-line, often retail products which have a downloadable trial version. Since the desired quality of the output video or audio file depends on the quality of the original material, it is recommended to select a retail product based upon satisfactory results with the trial version of the software.

Examples of retail software with trial downloads include the Xilisoft HD Video Converter (http://www.xilisoft.com/hd-video-converter.html) or Audio Converter (http://www.xilisoft.com/audio-converter.html); Blaze Media Pro (http://www.blazemp.com/); AVS Media tools (http://www.avsmedia.com/).

Alternatively, the open source software VLC VideoLAN (http://www.videolan.org/), designed to support a large number of multimedia files without the use of additional codecs, is capable of file format conversions.

Freeware for video/audio format conversions can be found on the Internet, but some freeware programs might be prone to crash during conversions, and some downloads may contain spyware or viruses. If a freeware product will be chosen for use, we advise to search the Internet for user reviews first.

B.7. Spreadsheets

XLS, XLSX => PDF/A and ODS

Like word processor documents, Office 2007 can save an Excel spreadsheet file as a PDF/A. Likewise, a spreadsheet file (or batch of files) can be printed to the Adobe PDF printer just like documents. Please refer to the section on **Word Processor Documents** above.

Spreadsheets are saved to ODS if they are saved in Open Office. This can also be done in Microsoft Office 2007 with Service Pack 2, as with ODT (see **Word Processor Documents)**.

B.8. Databases

DBF, MDB, ACCDB => CSV

This conversion is done by the data archive.

DBF files can be imported into Microsoft Access. Access tables and imported tables can be exported from the MDB or ACCDB file to CSV with an export wizard. Alternatively, a script can be written in Access to export tables using a form. We recommend writing a script which covers all data types, as CSV exports directly from the software (using Wizards or 'save as' functions) may be in danger of data loss: please refer to **Chapter 5.3.4: CSV export problems and directions for solutions**.

A form can be copied from one Access file into another. We suggest that one CSV export form should be created inside an empty database, this CSV-export can then be copied from the empty Access file to use for all table exports of MDB and DBF files alike. This ensures that each export will follow the same script.

Be sure to verify each of the individual tables in a database and omit empty tables in the export tables selection.

B.9. Statistical Data

SPSS, SAS, Stata => ASCII (archival format)

This conversion is done by the data archive.

The Interuniversity Consortium for Political and Social Research (ICPSR) developed an automated batch data conversion system which is currently used to convert SPSS data to ASCII text. The system operates from the command line in a Linux environment. Variable-level metadata are written out using the SPSS/Python plug-in, which allows direct access to the metadata at its source in the data file rather than writing out and parsing a data dictionary. SPSS/Python is also used to perform the ASCII data export. A public version of the SPSS/Python code including guidelines is available at http://sda.berkeley.edu/manh/makeddlsps.htm.

The ASCII export files could subsequently be converted to Unicode: please refer to the section on **Plain Text** above.

If the original data contained characters outside of the ASCII range, those characters will already be lost during the SPSS/Python export. It would be ideal to export statistical data directly to Unicode, but unfortunately there are no specialized tools for automated batch exports to Unicode at present, and not all software for statistical data can export to Unicode to begin with. SAS data can be exported to Unicode (http://support.sas.com/kb/13/666.html), support for Unicode in SPSS has recently been implemented and Stat/Transfer will support Unicode in the next release. Unicode support is for statistical data is altogether in development and in need of more research and tools.

B.10. GIS

TAB, SHP => MID/MIF

The TAB format is a MapInfo software proprietary format, the SHP Shapefile format is proprietary to ESRI software (ArcGIS). The formats can be imported in other GIS applications, and ArcGIS is capable of handling TAB-files and SHP-files can be imported in recent versions of MapInfo (9.5 onwards), but the use of different software may limit importing, editing and exporting options.

A GIS-file can be exported to the MID/MIF interchange format directly from MapInfo, using the 'Table Export' function.

Unofficial MapBasic programs for batch file utilities are found in the batchtools package by Justin Hyland, presently available at http://www.directionsmag.com/files/index.php/view/104.

Exporting a Shapefile to MID/MIF from ArcGIS requires the Data Interoperability extension, an official ESRI software extension. The Data Interoperability extension allows for the import of and the export to a wide variety of GIS-formats. It can be used for bulk processing: Data Interoperability tools including Quick Import and Quick Export will be added to the ArcToolbox.

Older Shapefiles sometimes have empty records in their associated DBF tables, leading to a mismatch with the number of shapes in the shapefile. Recent versions of ArcGIS will not be able to properly export these files and will give the error message "*Number of shapes does not match number of table records*" unless the excess lines are removed from the table. The files can be repaired with an unofficial tool called ShapeChecker, created by Andrew Williamson. The tool is presently available at http://www.geocities.com/SiliconValley/Haven/2295/.

B.11. CAD

DWG => DXF R12

Most CAD and GIS software will be able to save files to the DXF format. CAD formats do not have metadata tables so exporting a GIS file to DXF will result in the loss of all metadata. Export of files to DXF should therefore be limited to CAD files, the main format being DWG.

Not all GIS software can import the DWG format: only recent versions (9.5+) of MapInfo can import DWG, ArcGIS has capabilities of importing and exporting DWG. There are options to use GIS software but it is easier to directly save a DWG to DXF using CAD software. AutoCAD has the preferred DXF R12 format as file type option in the 'Save as' menu.

Batch conversions of DWG to DXF require specific software. One such utility is the Any DWG to DXF Converter, which can be purchased at http://anydwg.com/dwg-dxf/.

Appendix C. - File Format Identification Tools

If a digital archive will specify its preferred formats, there may be a desire to check which formats are already in storage. Which files in the archive are preferred formats? Which files may need to be converted? What sort of files does the archive consist of?

Another issue is the fact that files are not always identifiable from their extension.

C.1. Characterization tools

C.1.1. DROID – Digital Record Object Identification

http://droid.sourceforge.net/wiki/index.php/Introduction

DROID is a software tool to perform batch format identifications, using the PRONOM registry: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

Both DROID and PRONOM are developed by the UK National Archives.

DROID identifies files by internal and external signatures. The results can be output in an XML or CSV file. Files can get a positive identification, based on the binary signature of the file. Positive identifications can be specific, if a single format is identified, or generic, if several file formats contain the same binary signature.

If no binary signature information was available, files can get a tentative identification based on the file extension; this will list all possibilities of files with the same extension as registered in PRONOM. A file can also remain unidentified if there are no identifiers in PRONOM or if an error occurred during the identification.

DROID can be very useful to obtain an overview of file formats in an archive but it has its limits due to the PRONOM registry being incomplete. For example, PRONOM is presently lacking in information on GIS file formats and even erroneously ascribes the few Mapinfo formats it recognizes to the ESRI company.

The DANS archive has been scanned using DROID. A relatively low amount of files was positively identified. Tentative results often included the right identification, but a need remained for manual elaboration and correction of the results. The limits of the PRONOM registry, for example, caused any Mapinfo Interface Drawing format (MID) to be tentatively identified as a MIDI soundfile.

Additions to the PRONOM registry can however be submitted by anyone: http://www.nationalarchives.gov.uk/PRONOM/submitinfo.htm. A data archive can choose to provide PRONOM with any missing information on file formats that remain unidentified.

C.1.2. JHOVE – JSTOR/Harvard Object Validation Environment

http://hul.harvard.edu/jhove/

The JHOVE project is designed to perform format-specific identification, validation and characterization of files. It can be implemented as a Java application and it can also be invoked with a command-line interface or with a GUI interface.

JHOVE has modules for identification and validation of the following formats: AIFF, ASCII, Bytestream, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, UTF8, WAVE and XML. JHOVE can be used to obtain version-specific information, for example TIFF 6.0, HTML 4.0 and others.

JHOVE cannot identify any file outside of the above formats, like, for example, Microsoft Office formats (DOC, XLS, MDB).

The DANS archiving system EASY uses an implementation of JHOVE to automatically assign general technical metadata, such as format type and file size, to uploaded files.

JHOVE2 has been in development since late 2008, by a collaboration consisting of the California Digital Library, Portico, and Stanford University with the assistance of an advisory board comprised of members of several international institutions (archives and libraries), projects and vendors. JHOVE2 will severely improve JHOVE's characterization capabilities. Among other features, JHOVE2 will characterize files based on four specific aspects: signature-based identification, feature extraction, validation and rules-based assessment.

C.1.3. TrID File Identifier

http://mark0.net/soft-trid-e.html

TrID is a utility designed to identify file types from their binary signatures. It's database currently consists of 3773 file types, additional file types can be scanned for binary signatures and added to the database.

TrID can give partial matches for file scans. One test result for a JPEG image, for example, resulted in a match of 50.0% JFIF JPEG Bitmap, 37.5% JPEG BITMAP and 12.5% MP3 audio. Indefinite results such as these limit TrID's use for properly identifying files in an archive's collection, but TrID results may provide a useful lead for identifying obscure files or files with missing extensions.

The TrID download comes with the complete database in file-specific individual XML files. A TrID scan will refer to the matching XML, however not to the specific string match within the XML. A TrID XML file contains information on its creator but it holds no information on how it was compiled; there is no control on the quality. The scan results are therefore somewhat relative and subjective.

TrID is primarily designed for scanning binary files. Some (ASCII) text-based formats may still be recognized.

An on-line version of TrID exists at http://mark0.net/onlinetrid.aspx.

The present TrID database does not provide scans with information on specific versions such as PDF/A.

C.1.4. UNIX file command

http://www.darwinsys.com/file/ (homepage)

http://linux.about.com/library/cmd/blcmdl1_file.htm (manual page documenting the older version 3.39)

File is a standard UNIX program for determining the type of data contained in a file. It is open source, ships with every free operating system (Linux, BSD) and has been ported to other systems including Microsoft Windows and DOS.

The file command is a command-line tool that looks at the file's actual contents (instead of the extension) and reports what kind of data it contains if a match is found. The command will submit a file to three sets of tests:

- a file system test, which examines the return from a 'stat' system call to see if the file is empty or a special kind of file
- a magic number test, which checks for binary data strings in fixed formats, using information from a compiled magic file – or, if a file does not match any of the entries in the magic file, it is examined to see if it seems to be a text file
- a language test, which determines in what language a file is written by looking for particular strings and keywords.

C.1.5. ExifTool

http://www.sno.phy.queensu.ca/~phil/exiftool/

ExifTool by Phil Harvey is free platform-independent software designed for reading, writing and manipulating image, audio and video metadata (Exif: Exchangaeable image file format; a specification for images used by digital cameras, using existing JPEG, TIFF 6.0 and WAV formats with the addition of specific metadata tags), but capable of reading more types of formats such as PDF and XLS – a full list of formats which ExifTool can read can be found on the download page, along with instructions on how to use the tool.

C.1.6. Metadata Extraction tool

http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool

The Metadata Extraction Tool was developed by the National Library of New Zealand in 2003 to extract metadata from a number of file formats for preservation purposes. The software was released as open source in 2007.

Formats are limited to: BMP, GIF, JPEG, TIFF, MS Word (versions 2 and 6), Word Perfect, Open Office (version 1), MS Works, MS Excel, MS PowerPoint, PDF, WAV, MP3, HTML, XML. Scans for other file types will extract generic data recognized by the system (size, filename, date created).

The tool has a UNIX command line interface and a Microsoft Windows interface.

C.2. Format checkers

C.2.1. PDF/A checkers

The preferred format for several types of files including word processor documents is PDF/A, a subset of PDF designed for long-term archiving. Converting a file to a PDF does not automatically create a PDF/A.

The company PDFlib GmbH (http://www.pdflib.com/pdflib-gmbh/), which develops and sells development tools for server-centric generation and processing of PDF documents, recently published an extensive report on PDF/A validation tools: http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-04-03-Bavaria-report-on-PDFA-validation-accuracy.pdf

The chapter on **Conversion Tools** will provide some suggestions for tools with which to convert a file to PDF/A; here we will cover some tools with which to check if a file is PDF/A compliant:

C.2.1.1. Adobe Acrobat

http://www.adobe.com/products/acrobat/

A batch validation option was implemented in Adobe Acrobat 9.0 and improved in Adobe Acrobat 9.1. However, PDF/A includes certain references which should be met for PDF/A conformance (the specification XMP 2004 and the PDF 1.4 standard); while the Adobe Acrobat check is mostly thorough and accurate it does not fully check for these references.

C.2.1.2. Callas pdfaPilot

http://www.callassoftware.com/callas/doku.php/en:products:pdfapilot

Results of checks with pdfaPilot are very accurate and reliable, although it shares Adobe Acrobat’s issue of not fully checking certain references.

C.2.1.3. LuraTech LuraDocument PDF/A Validator

https://www.luratech.com/products/document-conversion-solutions/luradocument-pdfa.html

A command line tool designed to verify if a PDF meets the PDF/A standards.

We do not have any information on this tool’s capabilities of checking for the certain references which Callas and Adobe do not fully cover.

C.2.2. Image Checker

http://www.imagemagick.org/script/identify.php

The ImageMagick software suite is designed to create, compose and edit images in a variety of formats, either from the command line or accessed via the interface of a programming language. It has an option to ‘identify’ an image for its details. This feature will not check for specific versions of a file format (TIFF 1.0, 2.0, …), but it will list the image number, the file name, the width and height of the image, whether the image is colourmapped or not, the number of colours in the image, the number of bytes in the image, the format of the image. It will list many additional details with the ‘identify –verbose’ command.

ImageMacick also reports if an image is incomplete or corrupt.