



The taxonomy of data types and file formats in the AHDS collection

Final version

prepared by

Raivo Ruusalepp
Estonian Business Archives, Ltd.

October 2002

Table of Contents

TABLE OF CONTENTS	1
1. INTRODUCTION AND SCOPE OF THE REPORT.....	2
2. THE NATURE AND REMIT OF THE AHDS COLLECTIONS	3
3. A TAXONOMY OF DATA TYPES	9
4. THE TAXONOMY OF AHDS FILE FORMATS	14
4.1 TEXTUAL DATA	14
4.2 DATABASES.....	19
4.3 STATISTICAL AND SPREADSHEET DATA.....	23
4.4 DIGITAL IMAGES	26
4.5 DIGITAL VIDEO.....	31
4.6 DIGITAL AUDIO	34
4.7 CAD.....	37
4.8 GIS.....	40
4.9 OTHER FILE FORMATS.....	43
4.10 CONCLUSIONS	46
5. ASSESSMENT OF PRESERVATION MANAGEMENT PRACTICES	47
5.1 ACQUISITION FILE FORMATS.....	48
5.2 PRESERVATION FILE FORMATS.....	48
5.3 CREATION OF THE PRESERVATION VERSION	48
5.4 VALIDATION OF THE PRESERVATION VERSION.....	49
5.5 COMPRESSION TOOLS.....	50
5.6 STORAGE MANAGEMENT	50
5.7 AUTHENTICATION METHODS.....	51
5.8 PRESERVATION METADATA.....	51
5.9 CREATION OF THE DISSEMINATION VERSION.....	52
5.10 TECHNOLOGY WATCH.....	52
5.11 CONCLUSIONS.....	53
6. RECOMMENDATIONS	54
GENERAL RECOMMENDATIONS.....	54
DETAILED RECOMMENDATIONS	54
RECOMMENDATIONS FOR EXISTING POLICY TEXTS.....	55
APPENDIX I FILE FORMAT SURVEY LETTER AND QUESTIONNAIRE FORM.....	57
APPENDIX II BIBLIOGRAPHY AND USEFUL SOURCES OF INFORMATION.....	63
APPENDIX III A SELECTION OF STANDARDS AND DEFINITIONS OF FILE FORMATS	65

1. Introduction and scope of the report

Digital information is created, collected and stored in a wide variety of proprietary and standard formats. File format continue to evolve, becoming more complex as revised software versions add new features or functionality. It is not uncommon for software enhancements to “orphan”, or leave unreadable, files generated by earlier versions. The threat to ageing digital information has surpassed the danger of unstable media or obsolete hardware. The most pressing problems confronting managers of digital collections are data format and software obsolescence.¹

Digital preservation activities need to be planned and carried out on the file level where the file format the data has been saved in plays a crucial role. Type of the data provides a general set of criteria that need to be considered when performing digital preservation processes (e.g., conversion), but each file format has its specific requirements and functionality that may need to be retained throughout the preservation period. Therefore, the individual file formats need to be considered alongside the data types that are included in the AHDS collections.

The purpose of this report is to present an overview of data types and file formats in the collections of the AHDS Service Providers and assess risks associated with the file formats in use and preservation practices of file processing. The report is based on:

- desk research;
- questionnaire survey conducted among the AHDS Service Providers (see Appendix I);
- interviews with selected Service Providers staff;
- e-mail correspondence with individual members of Service Providers’ staff.

The report is divided into three sections:

- 1) analysis of the AHDS collection policy and remit for collection management,
- 2) analysis of file formats used for preservation for different data types, and
- 3) assessment of risks associated with current preservation practices of AHDS Service Providers.

The report concludes with recommendations for further action and improvement of preservation practices. Three appendices include the file format survey questionnaire form, list of useful information resources regarding the preservation of particular file formats and data types, and a list of standards and file format specifications referred to in the report.

Any comments and questions regarding the report should be addressed to Raivo Ruusalepp at raivo@eba.ee.

¹ G. Lawrence et al., “Risk Management of Digital Information: A File Format Investigation”, 2000, p. 1

2. The nature and remit of the AHDS collections

2.1 Through its Service Providers, the AHDS collects and manages high-quality digital resources that are of long-term interest and use to those researching and teaching in humanities disciplines. The AHDS is a geographically distributed service comprising a managing Executive and a number of Service Providers devoted to archaeology, history, literary, linguistic and other textual studies, the visual arts, and the performing arts. The Service Providers collect, preserve, catalogue, and distribute digital resources which are relevant to their subject areas.

The AHDS collections aim to comprise the full range of data types as necessary to serve the interests and needs of the user groups and include, for example, archaeological excavation archives, historical, reference, and other databases, electronic texts and musical scores, linguistic corpora, geographical information systems, image banks, digital sound and video, mixed media installations, etc. The collected data resources are both static (i.e., a coherent collection of digital information in its final state and unlikely to be changed by its creators) and dynamic (i.e., digital information being actively edited, updated or otherwise amended by the creator).

2.2 In order to fulfil its mission — to collect, preserve and disseminate — efficiently, the AHDS must ensure that all digital resources it acquires can be managed cost-effectively and retained accessible over the long term. The long-term management and preservation of a digital resource requires significant efforts on behalf of the AHDS to ensure that a particular resource remains accessible to, and reusable by, the intended user community. It is, therefore, important that the AHDS Service Providers assess the digital environment and documentation accompanying a digital resource when considering it as a deposit. Compared to archive services that have a statutory requirement to archive and preserve records, the AHDS has a more limited remit for setting requirements to the data and file formats that are deposited with the AHDS Service Providers. Nonetheless, the AHDS is influencing and controlling the state of resources it acceptions on at least three levels:

- Guidance and awareness raising for data creators and depositors;
- Requirements set in collection policy and depositor guidelines;
- Data structure, format and accompanying documentation form one set of evaluation criteria for acquisitions.

One of the reasons for the AHDS rejecting a data resource for preservation is if the resource cannot be used without proprietary or obsolete software.²

2.3 The AHDS and its Service Providers have produced a number of guides to good practice that are explicitly directed towards the data creators, data collectors and digitisation projects. One of the main aspects covered in these guides is the use of open, standardised file formats for storing data and good documentation practices.³

² “Managing Digital Collections: AHDS Policies, Standards and Practices”, Ch. 2.4.1

³ Guides to Good Practice in the Creation and Use of Digital Resources (<http://www.ahds.ac.uk/guides.htm>); cf. also “Creating A Viable Data Resource”, “Digitisation: A Project Planning Checklist”, “A Strategic Policy Framework for Creating and Preserving Digital Collections”

2.4 AHDS Service Providers' Collection Policies and Depositor Guidelines provide detailed recommendations regarding accepted data types, preferred file formats, and transfer media for deposited resources that ensure lower management, preservation, and dissemination costs and a higher potential for re-use and long-term preservation. Whilst such considerations shall usually not preclude the AHDS Service Providers from accessioning resources which have not been prepared in accordance with these recommendations, they will be taken into account when evaluating a resource for accessioning.

2.5 Another look at the data and file formats hosted by the AHDS is through the definition of five levels of collections that the AHDS has adopted. Some of these levels include the commitment to preserve the deposited data resource and some do not:⁴

- *archived*
The data resource is archived by the AHDS and the AHDS intends to preserve and keep the intellectual content of the resource available on a long-term basis. The resource will also normally be disseminated by the AHDS unless special arrangements have been agreed with a depositor e.g., to restrict access for a specified period of time.
- *served*
The data resource is accessioned, catalogued and disseminated by the AHDS but another institution has primary responsibility for content, maintenance and long-term preservation. This collection level may include 'mirrored' resources where a copy of a digital resource residing elsewhere is hosted by the AHDS to improve access, or resources held, maintained, or preserved by collaborating and commercial agencies, which are licensed and disseminated by the AHDS.
- *brokered*
The data resource is physically hosted elsewhere and maintained by another institution but the AHDS has negotiated access to it with a collaborating agency and includes metadata and links for the resource in its catalogue, or AHDS users are able to locate and cross-search, and in some circumstances acquire access to it.
- *linked*
The data resource is hosted elsewhere and the AHDS provides a web link pointing to it at that location from its webpages. The AHDS has not accessioned that resource or negotiated a collaboration agreement with the agency which maintains it and has no control over the information or formal agreements for access to it.
- *finding aids*
Electronic finding aids and metadata held by the AHDS which will facilitate discovery and searching of digital resources. This metadata is associated with digital resources such as collections at the AHDS or elsewhere but may be stored, managed and maintained separately from them.

It is only the first category of accessions — archived — that the AHDS Service Providers take a full retention responsibility for. With the served and brokered accessions, the AHDS may hold dissemination versions of a data resource, but has no primary responsibility for the preservation of the resource. These two categories have also been defined as 'volatile data resources',⁵ where the creator or manager/maintainer is actively editing the data resource and the AHDS acts as an access point to the resource. An exception to this rule is where the data resource creator or manager lacks network access but wishes to make a volatile data resource

⁴ "Managing Digital Collections: AHDS Policies, Standards and Practices", Ch. 2.3.2

⁵ ADS Collections Policy, 2.1.4; OTA Collections Policy, 2.1.4

available for re-use: the AHDS Service Provider may then take a “snapshot” of the resource, which is subsequently periodically refreshed.

The last two categories form the so called 'virtual accessions'⁶ where the AHDS Service Provider does not take physical custody of the data resource, but only points to it or holds metadata about the data resource in its catalogue.

2.6 Resources offered for deposit will be evaluated to see how (or whether) they can viably be managed, preserved, and distributed to potential secondary users. Factors influencing this evaluation include:⁷

- documentation: resources should be supplied with appropriate and sufficient documentation to satisfy the requirements for reuse by members of the academic community and management of the resource by AHDS.
- Intellectual Property Rights and licensing. Resources should be accompanied by a signed Common Deposit Form providing the rights needed by AHDS to manage and distribute the resource
- Quality of the digital environment. The digital environment of the resource will be assessed for its integrity, ease of use, and portability to different technical environments.
- Long-term preservation. The standards and formats used to create the resource should enable and not preclude long-term management and re-use. The data is of a type with which the AHDS has expertise in managing or may easily obtain expertise and expert advice.
- Cost. Resources required to accession; enhance or create documentation; validate, migrate, or re-format; manage and preserve; and provide access to any potential acquisition.

Although the use of obsolete and proprietary file formats is listed as a sufficient reason for the AHDS to reject a data resource for accession, most Service Providers do not make it a definite rule and are prepared to consider the data resource. A summary of statements on file formats as the evaluation criteria of data resources, based on the Service Provider Collection Policies, is presented in Table 1.

Reference	Statement
ADS	
2.3.2.3 Suitability for Digital Preservation	If the format in which a dataset is stored means that the digital resource is irrecoverably obsolete upon presentation to the ADS this will be sufficient reason for recommending that the dataset not be accessioned.
HDS	
	<i>No criteria set relating to file formats.</i>
OTA	
2.3 Criteria for Evaluating Electronic Datasets	Evaluate how (or whether) they may viably be managed, preserved, and distributed to potential secondary users. Factors influencing this evaluation include: Format: scholarly electronic texts or linguistic corpora whose

⁶ the VADS Collections Policy (Ch. 2.5 Acquisition methods) makes a distinction between 'Physical accession' that covers the archived level of accessions and 'Virtual accession' where data is not stored on the VADS server.

⁷ “Managing Digital Collections: AHDS Policies, Standards and Practices”, Ch. 2.4.4

	format is such as to enable their long-term management and reuse by members of the academic community. Documentation: scholarly electronic texts or linguistic corpora which are supplied with appropriate and sufficient documentation to satisfy the requirements for reuse by members of the academic community.
PADS	
Assess condition of resources and supporting documentation	If the format of the resource significantly impedes its use or the quality of documentation which accompanies it is such that the PADS cannot offer it as a working resource then the PADS may have to reject the materials offered.
VADS	
2.4 Quality of digital environment	VADS does not plan to refuse datasets on the grounds that best use of digital environments had not been made. One of the key aims of VADS is to apply Standards of best practice to accessioned data and promote the use of these standards in the Visual Arts Community. However if the format in which the resource was provided was irrecoverably obsolete for example or the images provided of a prohibitively low quality then this may be grounds for not recommending that a particular dataset was accessioned.

Table 1. File formats as evaluation criteria of potential deposits.

2.7 In accordance with their mission, the AHDS Service Providers have to preserve the deposited data resources in a usable format and provide simple and functional access to them. This is achieved by defining and creating three different versions of each of the deposited files that form a data collection:

- 1) *original version* of each file in the data collection as it is deposited with the Service Provider;
- 2) *preservation version(s)* of each file in the data collection held in format(s) which may differ from the original but are deemed suitable for long-term preservation of the intellectual content of the data resource;
- 3) *dissemination version(s)* of each file in the data collection with added value created by the Service Provider.

The original version of data resource is comprised of the originally deposited data files and the deposited documentation. All AHDS Service Providers keep the originally deposited files for comparison and validation for the created preservation and dissemination versions. In the long term it is envisaged that the originally deposited files may be migrated to new formats or emulation tools built to recreate the information content (and, if possible, the experience of using the data resource).

The preservation version is intended to guarantee the long-term survival of the deposited data resources. In creating the preservation version, the original files may be converted to new file formats, in order to preserve their information content, but the preservation format can also be identical with the original format (i.e., if the data resource was deposited in a file format suitable for long-term preservation).

The dissemination version is intended for using with currently available software. Dissemination file format may be different from the original and/or preservation formats, but does not have to be. Dissemination versions are created as a convenience for users and they do not need to be preserved for the long term.

2.8 The process of creating the preservation and dissemination versions may involve altering the original data resource. This is done in order to guarantee the retention of an accessible resource that is easy to distribute and use. The preservation version should:⁸

- Minimise the need for future migration of files;
- Minimise dependency on proprietary file formats;
- Maximise software and hardware independence;
- Ensure that the documentation supplied with the data is preserved.

The creation of a preservation version may mean complex data structures being decomposed into a number of simpler data structures; proprietary file formats being changed to open file formats; documentation being edited, created, or digitised; and creation of new metadata.

Although the preservation version is often suitable for dissemination to users, it is not always the case. Creating a preservation version may involve discarding useful functionality, or using file formats that are not automatically compatible with common software programs. When these disadvantages are significant, a separate dissemination version of the data resource that is designed to be used with currently available software can be created.

The paper documentation that is sometimes submitted with the deposited resource is usually scanned and stored together with the data resource in a format suitable for medium- to long-term digital preservation. The documentation files should be treated equally with textual data files for preservation purposes.

2.9 Each data resource that is preserved by the AHDS is, thus, comprised of four components: the original version, the preservation version, the dissemination version, and documentation. Each of these components includes files in up to three formats: original, preservation and dissemination formats, which may or may not overlap. The following analysis will use this three-tier distinction between the file formats and treat them all as equal objects of preservation activities. The analysis groups the file formats according to their data type and includes documentation of a data resource, considering it as a specific case of a 'data type'.

2.10 The following review of data types and file formats managed by the AHDS Service Providers will only cover the archived level of accessions since these are being both preserved and disseminated by the Service Providers. The scope of this study is, therefore, narrower than the list of data resources and file formats that can be accessed through the AHDS catalogue.

2.11 Conclusions

In building their collections, the AHDS Service Providers have no statutory obligation to accept any digital resources offered to them, but the AHDS has adopted an open and flexible collection policy in order to ensure survival of and access to a wide variety of scholarly resources. Although the nature of AHDS collection-building principles does not permit it to set stringent requirements to the data resources transferred to the Service Providers for

⁸ "HDS Collections Manual", D.2.2

preservation (as, for example, a national archive is required to do), the AHDS reserves a right to refuse the acquisition of a data resource if it is not compliant with the standards set by the AHDS. Among such evaluation criteria the file formats used for saving the data resource are among the top three. However, the AHDS provides guidance on good practice in data creation and documentation and is prepared to work with the data resource depositor to bring the resource to the level required for acquisition.

The AHDS does not take physical custody of all the data resources it provides access to: it may broker access to data resources hosted by other institutions and act as an access point to this data or provide references and links to data resources elsewhere.

Those data resources that are physically deposited with the AHDS Service Providers will be processed according to the AHDS collections management procedures and will continue their existence in up to three tiers: the original version, the preservation version, and the dissemination version. All three versions together, combined with the accompanying documentation, form one single data collection. The two new versions that are created from and in addition to the original deposited resource, serve the purpose of ensuring the preservation of accessibility for the long-term and simple use of the data resource for the short term.

Each deposited data resource may comprise several data types and multiple files in different formats. Documentation submitted with the data resource may be in paper form, but will be digitised in the course of ingest. If the deposited data resource includes files that are saved in formats that are not deemed suitable for long-term retention, they will be converted, in the process of creating the preservation version, into new formats that the AHDS has chosen for long-term preservation (usually open, standard file formats). Equally, if the original file formats and preservation formats are not suitable for easy dissemination and use of the data resource, or if their functionality is limited when used with current software programs, the AHDS will convert the files in the process of making the dissemination version into formats that are best suited for use with currently widespread software packages. Consequently, every data resource deposited with the AHDS will be stored in up to three copies and may be stored in several formats, thus reducing the risk of loss of access to the information contained in the file.

The three versions included in a data collection managed by the AHDS may differ, accordingly, on the level of file formats that they contain. If the deposited data resource used file formats that have been chosen for preservation and/or dissemination, no file conversion is undertaken during the creation of preservation and dissemination versions. The primary objective of file conversions performed by the AHDS is the preservation of access to the intellectual content of files and, where appropriate, less regard is paid to the preservation of properties of file formats that contribute to the 'look and feel' of their usage. Where possible and viable, these aspects may be preserved or described in the documentation of the data resource.

The documentation accompanying a data resource is stored digitally in file formats that are suitable for medium- to long-term preservation and can be disseminated in the same format.

The following chapter will review and analyse the data types and their file formats.

3. A taxonomy of data types

3.1 While in programming languages data types are defined as sets of data with predefined characteristics (e.g., integer, floating point unit number, character, string, pointer), on the application software level it is more common to use broader categories of data types: text, databases/tabular data, graphics images, digital audio and video, GIS, etc. The AHDS Service Providers have given a distribution of their collections across different data types in their Depositor Guidelines and Collections Policies. The two lists of data types that are accepted by the Service Providers differ in the two policy documents: the first list is following the more “orthodox” way of defining a data type, whereas in the second list the data types have often been defined through types of resources that the data depositors are likely to create, use and deposit. An overview of the two data type taxonomies is provided in Table 2 below.

While it is not difficult for the readers of these two policy texts to understand the meaning of each use of ‘data type’, it may be somewhat less confusing if a different definition was used in the Collections Policy, for example, ‘types of data resources’ (as it is defined in the PADS Collections Policy). See also Recommendations in chapter 6.

Data Type Definitions

ADS		HDS		OTA		PADS	VADS	
Dep. Guide Section E	Coll. Pol. 2.2	Dep. Guide ⁹	Coll. Pol. ¹⁰	Dep. Guide 9.5	Coll. Pol.	Dep. Guide & Coll. Pol. ¹¹	Dep. Guide 2.B	Coll. Pol. 2.3
Texts	Electronic texts	Text	Alphanumeric	Plain text	Electronic texts	Music and dance notations	Text	Electronic text
	Electronic journals	Textbases	Texts	Rich Text Format	Textbases	Electronic catalogues		Electronic text containing images
	Bibliographic finding aids	Electronic texts		Modifiable binary texts	Corpora			
				Unmodifiable (binary) texts				
Databases	Databases	Databases	(Numeric)		Databases	Databases (Performance history, music manuscript)	Database	Database, multimedia
			(Alphanumeric)					Database, text
Spreadsheets		Spreadsheets	Numeric				Spreadsheet	
Statistics		Statistical package	Numeric					
Images	Aerial photographs	Scanned images	Digitised images		Digital image data	Digitised photographs	Image	Image set
		Bitmap images			Image based documents	Still images		
		Vector images						
Movies		Moving images			Moving image	Digitised moving images	Moving image	
						Video arts		

⁹ <http://hds.essex.ac.uk/depguide.asp#depositing>; What is Involved in Depositing Data?, What is the Scope of the History Data Service Collection?

¹⁰ <http://hds.essex.ac.uk/collpol.asp#scope>; Scope of Collections

¹¹ “Types of resources accepted” — descriptions of data types in both PADS policies match

ADS		HDS		OTA		PADS		VADS	
		Sound			Audio	Digitised sound			
GIS	Topographic survey		Digitised boundary data						
	GIS								
CAD	Building survey					CAD-CAM resources	CAD		
							3D		
					Hypermedia	Web resources			
Virtual Reality	Visualisation				Virtual reality model		Virtual reality		
Geophysics	Geophysics data								
					(Applications software)	Courseware	Multimedia application	Informational software	
						Bespoke teaching packages	Applications	Application software	
							Desktop publishing		
Undesirable or unlikely deposit data types									
	Applications software				Applications software				
	Teaching materials				Digitised audio material				

Table 2. Definitions of data types accepted by the AHDS Service Providers.

3.2 The questionnaire survey of AHDS Service Providers identified the following distribution of data types in the Service Provider collections (see Table 3 below). Data volume figures are in Gigabytes and are rounded to two decimal points. The volume of each data type in the AHDS holdings includes all three versions of data resources that are being preserved.

Data Type	ADS		HDS		OTA		PADS		VADS		Total
	Has	Volume	Has	Volume	Has	Volume	Has	Volume	Has	Volume	
Text	✓	4.18	✓	3.06	✓	2.08	✓	1	✓	0.1	10.42
Database	✓	12.85	✓	40.78			✓	1.32	✓	0.5	55.45
Spreadsheet	✓	0.52	✓	0.77	✓	0.001					1.291
Image	✓	29.98	✓	6.10			✓	3	✓	500	539.08
Video	✓	0.68					✓		✓	0.5	1.18
Audio					✓		✓	0.18			0.18
CAD	✓	0.79									0.79
GIS	✓	0.27									0.27
Other	✓	1.11							✓	0.13	1.24
Total:		50.38		50.71		2.081		5.5		501.23	609.90

Table 3. Data types and volumes held by the AHDS Service Providers.

Comparison of data types defined as acceptable to deposit and data types that the AHDS Service Providers actually hold demonstrates that:

- 1) the Service Providers are prepared to accession a wider range of data types than they have received so far;
- 2) adhering to the principle that 'OTA will accept texts in most formats as long as they are accompanied by a suitable level of documentation' (OTA Depositor Guidelines, Ch. 9.5), OTA has accessioned a spreadsheet data resource, (OTA has not defined spreadsheets or statistical data as their deposit data types, cf. Table 2 above).

3.3 The total volume of data in the AHDS collections, compared with a survey conducted little over a year earlier, shows an increase for all Service Providers, in particular for the VADS and PADS (see Figure 2 below). In terms of volume of data types, the digital images and textual data have increased most in the AHDS collections over the past year and digital video has appeared as a new data type for preservation (see Figure 1 below).

3.4 Conclusions

The AHDS collection is spread widely across data types, but the difficult data types from the long-term digital preservation point of view form a rather small proportion of the overall volume of collections.

The next section will look at each data type and analyse the file formats used by the AHDS Service Providers to store and disseminate the data resources.

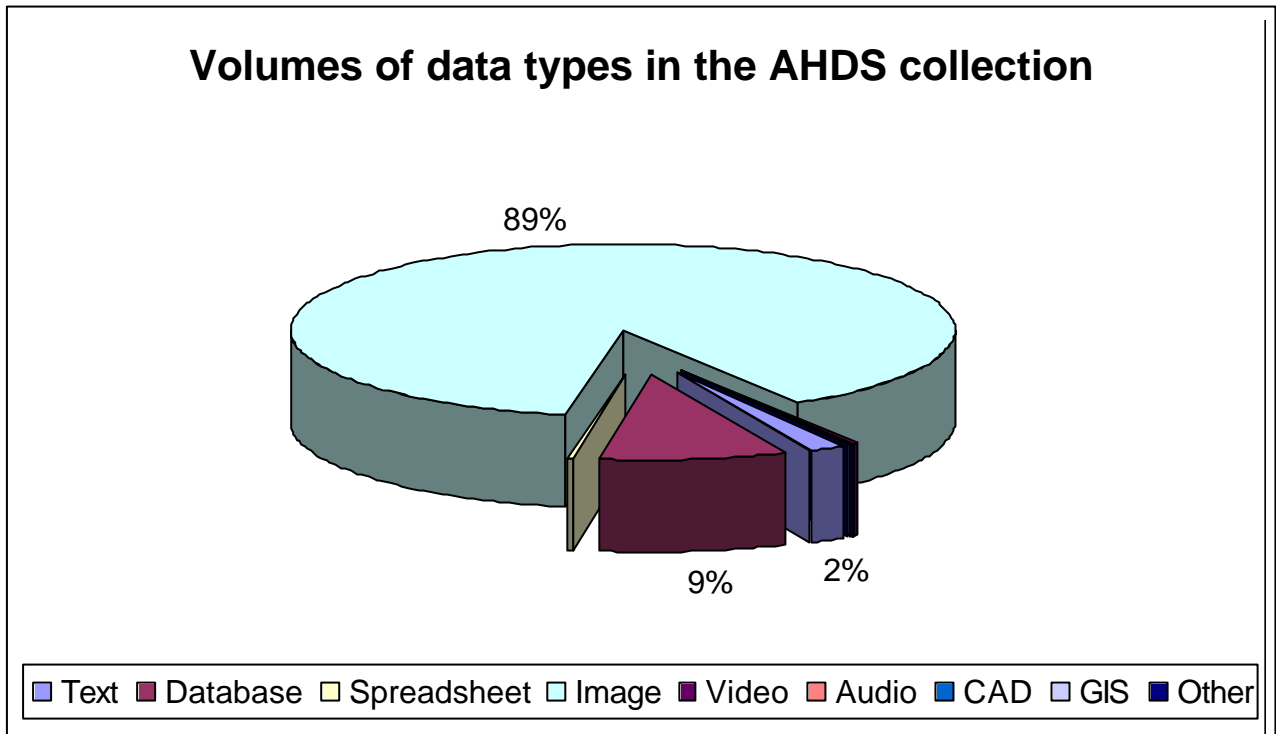


Figure 1. Distribution of data types in the AHDS collections.

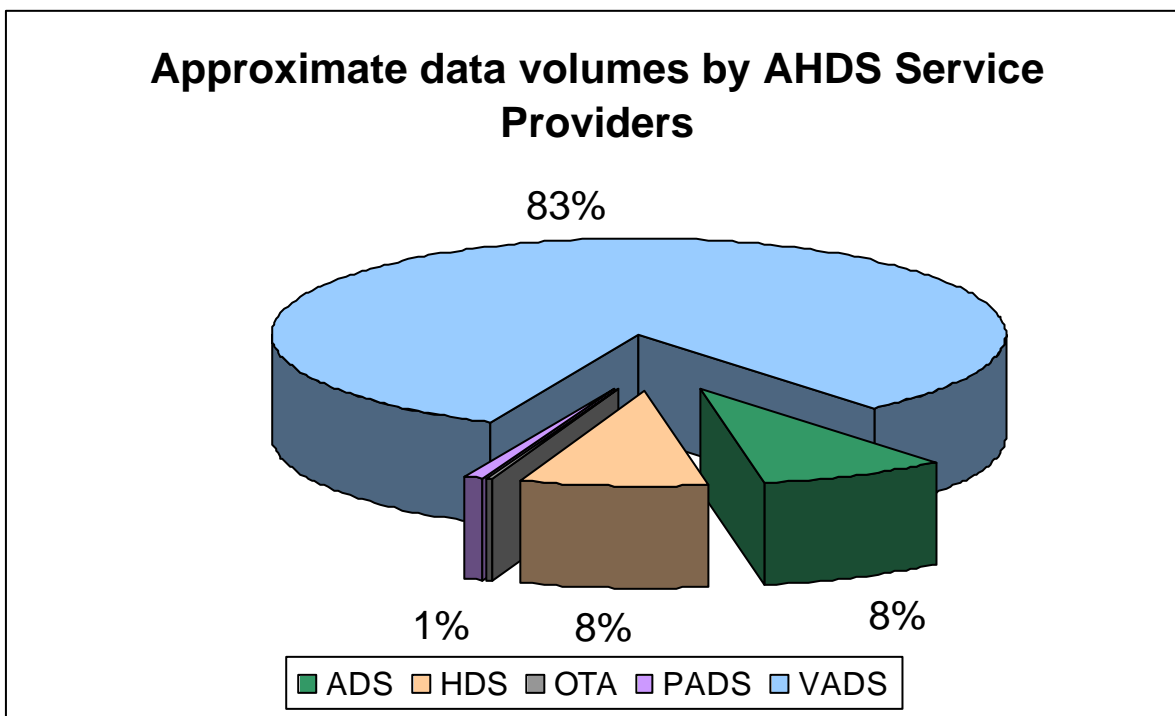


Figure 2. Approximate data volumes held by the individual AHDS Service Providers.

4. The taxonomy of AHDS file formats

The following section will analyse the file formats accepted and created by the AHDS Service Providers for long-term retention and dissemination to their users of the archived data resources.

The Service Providers have listed preferred file formats that they are prepared to accept the data resources for accession. They have also defined certain file formats as suitable for long-term preservation and dissemination. However, not all of these file formats actually exist in their collections as data has been deposited only in a limited number of formats. A survey and analysis of these will follow.

4.1 TEXTUAL DATA

4.1.1 Data resources that are created by using a text editor or word processing software are commonly saved in one of three broad types of file formats:

- As plain text files that contain just the text as a sequence of characters;
- As a structured or marked-up plain text file that is independent of the hardware platform;
- As a fully formatted text file that contains the characters, structure and layout but is not independent of the hard- and software used to read it.

The AHDS collections include all three types of text files.

4.1.2 The encoding schemes used for representing the characters and symbols in a text file can vary, but are, as a rule, well compatible with each other.¹² The 7-bit encoding scheme that was developed in the 1970s and was called American Standard Code for Information Interchange (ASCII), defined 128 symbols (ANSI INCITS 4-1986). An international standard has also been issued for this coding: ISO/IEC 646:1991 “ISO 7-bit coded character set for information interchange”. Subsequently IBM developed the Extended Binary Coded Decimal Interchange Code (EBCDIC) that supported 256 characters and some time later the extended, 8-bit ASCII code. The 8-bit encoding scheme became an international standard ISO/IEC 8859 which has been issued in 16 parts (latest revisions between 1998-2001) that define the structure for 8-bit coded character sets for non-English languages. The initiative of computer industry to develop a 16-bit encoding scheme, which they called Unicode, was later incorporated into another ISO standard ISO/IEC 10646:2000 “Universal Multiple-Octet Coded Character Set (UCS)” that comes in two parts (latest amendment is from 2002). The UCS and the Unicode (also referred to as UCS-2) are designed to allow most of the characters and symbols in use throughout the world to be coded. The first seven and eight bits of the Unicode/UCS standard still represent ASCII and the extended ASCII respectively. Another code — UTF — uses a variable number of bytes for coding symbols and is designed to make it easier to transfer multi-byte characters between different computers. It does this by avoiding the use of 8-bit control characters in the multi-byte character sequences.

The main disadvantage of the plain ASCII is that the 128 symbols cannot render languages other than Western European (and even those not well) and it cannot deal with alternate character sets or writing systems (e.g., right to left). This has led to a conclusion (in the past)

¹² cf. Ch. Dollar, “Authentic Electronic Records: Strategies for Long-Term Access”, 1999, Ch. 1.1.1

that the plain ASCII (the 7-bit that is also referred to as ‘unadorned’) is not suitable for the scholarly and preservation community, where the Unicode should be used instead.¹³

The preservation of plain text files (ASCII or Unicode) is straightforward and carries no inherent risks.

The Rich Text Format (RTF) was developed by Microsoft as an interchange format that retains all the formatting instructions of a document; it is a proprietary product developed and maintained by Microsoft, but its widespread use qualifies it as an *de facto* standard format that is suitable for medium- to long-term preservation of textual material.¹⁴

For a long period, Standard Generalised Mark-up Language (SGML) was considered the only “safe” format for long-term storage of complex textual data files, and it remains a safe format, as long as the structure of a file is described in a DTD and retained alongside the text file itself. The same applies for the Extensible Mark-up Language (XML) and other mark-up based formats. When SGML or XML are used as encapsulation tools, where a DTD is defined for an aggregation of documents that are not all textual files, care must be taken that all elements of such an ‘envelope’ are described and treated for preservation.

The Hypertext Mark-up Language (HTML), although plain text-based and therefore technically not difficult to preserve, is still not considered to be very stable for archival purposes (although a stabilised version has been defined by ISO/IEC 15445:2000), nor very suitable for long documents and therefore a low- to medium risk file format for long-term retention of textual data. The risks associated with this file format increase when the file includes links to objects outside the file which are considered to be part of the same document or data resource.

While the marked-up text files usually preserve the structure of the document, they may not always be able to retain the original presentation of the document and/or the more complex functionalities offered by word-processing packages.

The Portable Document File Format (PDF) is designed to replicate a document exactly as it appeared to the creator of the document. PDF is a platform-independent document format developed by Adobe as a follow-up to its Postscript language. Although the PDF specification is open and freely published,¹⁵ the format is maintained by the Adobe Systems Inc. who considers it the open *de facto* standard for electronic document distribution world-wide.¹⁶ PDF documents can be protected against editing or revising which makes it a safe format in one sense, but archivists do not recommend the PDF as a format for long-term preservation, although its wide, almost universal, use and published specification render it a medium-risk preservation file format. Several research projects are working on emulation tools for PDF software and converters from PDF to RTF and ASCII also exist.

4.1.3 The file formats used by the AHDS Service Providers for accessioning, preserving and disseminating textual data resources are listed in Table 4 below. The source data for the table were collected through a survey of file formats that each AHDS Service Provider filled in (see Survey questionnaire in Appendix I). Comments on file formats and volume of data follow after the table.

¹³ cf. J. Coleman, D. Willis, “SGML as a Framework for Digital Preservation and Access”, 1997, pp. 4-5

¹⁴ for levels of standards see DLM-Forum, “Guidelines on best practices for using electronic information”, 1997, pp. 50-51

¹⁵ <http://partners.adobe.com/asn/developer/acrosdk/docs.html#filefmtspecs> for version 1.4

¹⁶ <http://www.adobe.com/products/acrobat/adobepdf.html>

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Plain text						
plain text	1.75	plain text	0.4	plain text		2.15
Marked-up text						
SGML		SGML		SGML		
XML	0.25	XML		XML		0.25
SGML/TEI	0.25	SGML/TEI		SGML/TEI		0.25
HTML	0.38	HTML	0.8	HTML		1.18
LaTeX						
TeX						
Proprietary						
RTF	0.01	RTF	0.89	RTF		0.9
PDF	1.75	PDF	0.8	PDF	1.32	3.87
MS Word	0.03			MS Word (v. 97 <)	0.515	0.545
PostScript	0.01					0.01
WordPerfect	0.02					0.02
Idealist	0.07	Idealist				0.07
NUD*IST		NUD*IST				
Expected new formats in the near future						
XML						
Problematic formats for the Service Providers						
WordStar						

Table 4. File formats and data volumes for textual data in the AHDS collections.

NOTE: The volume of data archived in each format is a total of all Service Providers that keep the particular file format. The data volumes are given in Gigabytes and are approximate estimates. Where the preservation and dissemination format are identical or the data are deposited in a format suitable for preservation and dissemination, the calculation of the total volume of this format includes the data volume only once (i.e., if the total data volume deposited in SGML/TEI is 0.25 Gb, it is not tripled, since the preservation and dissemination formats are the same).

The first column lists file formats that the AHDS Service Providers are accepting deposited data in. The list of deposit file formats includes also those that the Service Providers are prepared to accept but thus far have not accessioned any data in these formats (where the data volume figure is missing). The preservation formats column lists formats that the Service Providers have chosen as safe for long-term preservation. Not all Service Providers use all the same file formats, but plain text and RTF are universal. The data volume figure for PDF format includes the documentation that forms part of the metadata for archived resources. The documentation files are objects of preservation equal to textual data resources, even though the purpose of their creation is somewhat different. The dissemination formats column lists file formats that the Service Providers are commonly making their textual data resources available to users and that are being preserved for short- to medium term. The total volume of textual data is divided between file formats as follows:

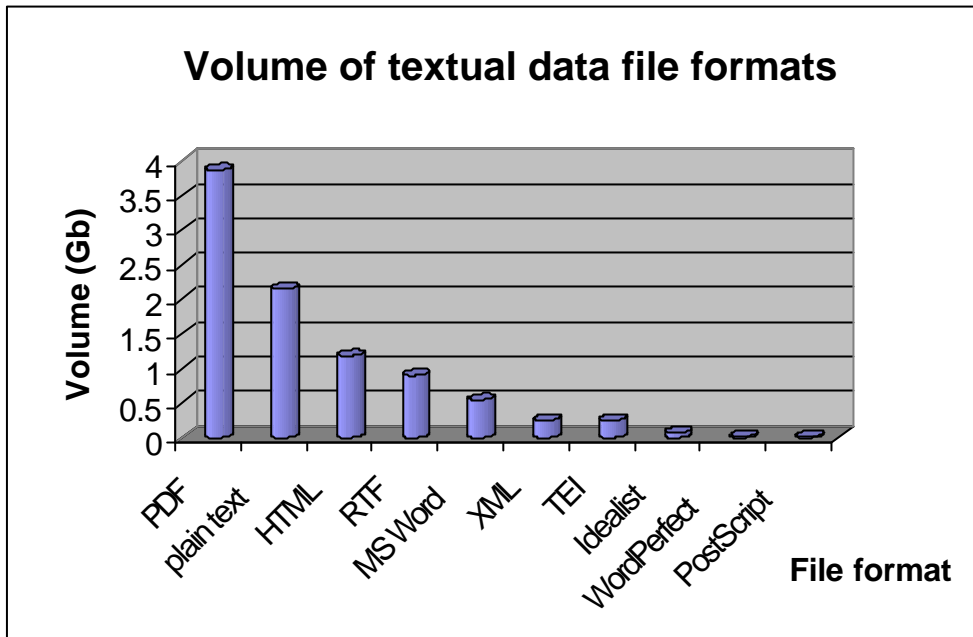


Figure 3. File formats for textual data in the AHDS collections.

The survey of file formats in the AHDS collections identified that the text encoding schemes that are being used are either ISO 8859 based, 7-bit ASCII (in one case) or UTF-8 compatible. Converting the already existing 7-bit ASCII data resources into new encoding schemes would yield little and is not recommended. However, raising the awareness of data creators and depositors about the usefulness of Unicode/UTF-8 would help to standardise the character codes in the AHDS collections in the future.

4.1.4 Risks and factors affecting the preservation of textual data file formats

Textual documents and data files are considered the simplest to preserve because they are well-scoped – (ideally) they contain all the information relating to the document within one file that is the object of preservation. Complexity of a textual file remains low if a standard mark-up language is used, but rises when a document links to other objects outside itself, when extra functionality for document formatting (e.g., footnotes, pagination, table of contents, etc.) is introduced, or the document contains a macro, is intended to work in a networked environment or contains HTML linkages.

The file formats chosen for long-term preservation by the AHDS Service Providers can be considered almost risk-free, as all formats are standard or at least semi-standard formats that are proprietary, but developed as ‘open formats’ (e.g., RTF, Idealist). Files in other text formats in the AHDS collections that are mostly stored for dissemination, can be converted into XML or PDF at little extra cost and therefore pose little risk of loss in the long term.

The file format survey included a question about the file formats that the AHDS Service Providers are expecting to have to deal with in the near future. The XML was mentioned twice as such format, but no indication was given whether any preparations are being made for accessioning and preserving these files. While XML has been praised for its flexibility and ease of use, it has not been thoroughly tested as a preservation format, yet, and the AHDS has an opportunity to make a contribution to the international best practice here.

Only one file format of textual data was mentioned as problematic, i.e. where access to the archived data resource may be lost soon. The questionnaire was not specific enough to identify whether the problem has been solved by creating a different preservation format from the originally deposited file format (WordStar) or whether urgent action is needed to save the file. As a minimum level of security measures, the files should be saved as plain text files. Alternatively, the Adobe PageMaker software can import WordStar files (versions 3.3 to 6) and save them as PDF, RTF or ASCII files.¹⁷

4.1.5 Conclusions and recommendations

The preservation file formats chosen by the AHDS Service Providers for storing textual data are standard and reasonably open formats that carry little risk of loss of access to the data.

All AHDS Service Providers should ensure that at least one mark-up language (e.g., XML) is listed as preferred preservation format aside the plain text and RTF which both have some limitations as preservation formats.

A Technology Watch service should be established to explore the ways of exporting the proprietary text formats into mark-up formats like XML. For example, the Acrobat has released a trial version tool for converting PDF files into XML.

¹⁷ <http://www.adobe.com/support/salesdocs/21ce.htm>

4.2 DATABASES

4.2.1 Database files are perhaps the type of electronic records that the archivists have had the most experience in preserving, yet there is still no easy way of retaining for the long-term the constituent elements of a database in their original functional capacity:

- the data;
- the structure of the database;
- the interface with the data (queries, forms, etc.).

Currently, no high-level standardised format exists for the long-term preservation of databases. The software programs used to create and manage databases are proprietary and not as platform-independent as are the data. However, the variety of commonly used database software packages do have some common characteristics that are useful for preservation purposes:¹⁸

- most database packages can export ASCII versions of tables that make up the database;
- most databases can generate, often in electronic form, documentation, such as definitions of the content of tables and the relationships between tables;
- it is possible to capture the content of data entry screens, often by using an image processing software package to create images of each data entry screen if the software cannot provide an image.

4.2.2 The best options for retaining access to database resources are:

- preserve the original file and software environment that created it (means no loss in data and functionality, but will become costly over long-term and therefore not a viable strategy);
- use a wide-spread proprietary format which can be re-read by several database systems (means some loss in functionality of access to the data may be lost, e.g., a data entry form);
- establish a conversion path or strategy for keeping the data accessible with current software (means, usually, losses in functionality and structure of the database);
- save data as plain text/flat files/data files and describe the structure and functionality separately (means loss of all functionality and structure).

While the last strategy is the safest from the point of view of preserving the data, it does not ensure the preservation of original functionality of using the file with a database management system and usually incurs high processing costs for creating the preservation format and user formats from it. Practically the only aspect of databases that has become non-proprietary and has been standardised is the access techniques. The Structured Query Language (SQL) became an ANSI standard in 1986 and later an international standard which has now reached its fourth version (ISO/IEC 9075-4:1999). SQL can define both the database data and methods of access to them.

4.2.3 The AHDS Service Providers accept database data in a variety of formats, but use a limited range of preservation formats. The results of the file format survey are presented in Table 5 below.

¹⁸ N. McGovern, "Cornell University Electronic Student Records Systems Project Report", 2000, Appendix B

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
Tab delimited text	0.01	Tab delimited text	1.84	Tab delimited text		1.85
Comma delimited text (.csv)		Delimited text		Delimited text		
MS Access	0.25	MS Access		MS Access		0.35
MS Access (<v.97)						
SQL Script		SQL Script		SQL Script		
				WebObjects	0.50	0.50
Oracle	12.10					12.10
FileMaker PRO	0.39					0.39
MySQL	0.10					0.10
DBase (.dbf)	0.01					
Paradox						
FoxPro (.dbf)						
Acceptable formats						
.dbf		.dbf	0.17	.dbf		0.18
.csv		.csv		.csv		
MS Access				MS Access (v.97<)	0.10	
Paradox				OpenBase		
FileMaker PRO						
Database structure description methods						
SQL						
Textual description						
Entity Relationship Models						
Not described						
Expected new formats in the near future						
-						
Problematic formats for the Service Providers						
FileMaker PRO						
Cardbox						

Table 5. Database file formats used by AHDS Service Providers.¹⁹

Table 5 lists file formats in which the AHDS Service Providers accept for deposit, and create for preservation and dissemination for database data. The aggregate data volume for each file format is approximate. The table does not currently include the 39 Gb of 1881 Census data that the HDS holds.

Different Service Providers have chosen different policies on some deposit file formats, hence, for example Paradox and MS Access appear as both preferred format and acceptable format. The data disseminated via WebObjects system is in a special file format which is intended for providing access. The overall distribution of database data between file formats is presented as Figure 4 below.

¹⁹ see Note under Table 4 for explanation of calculations in the table.

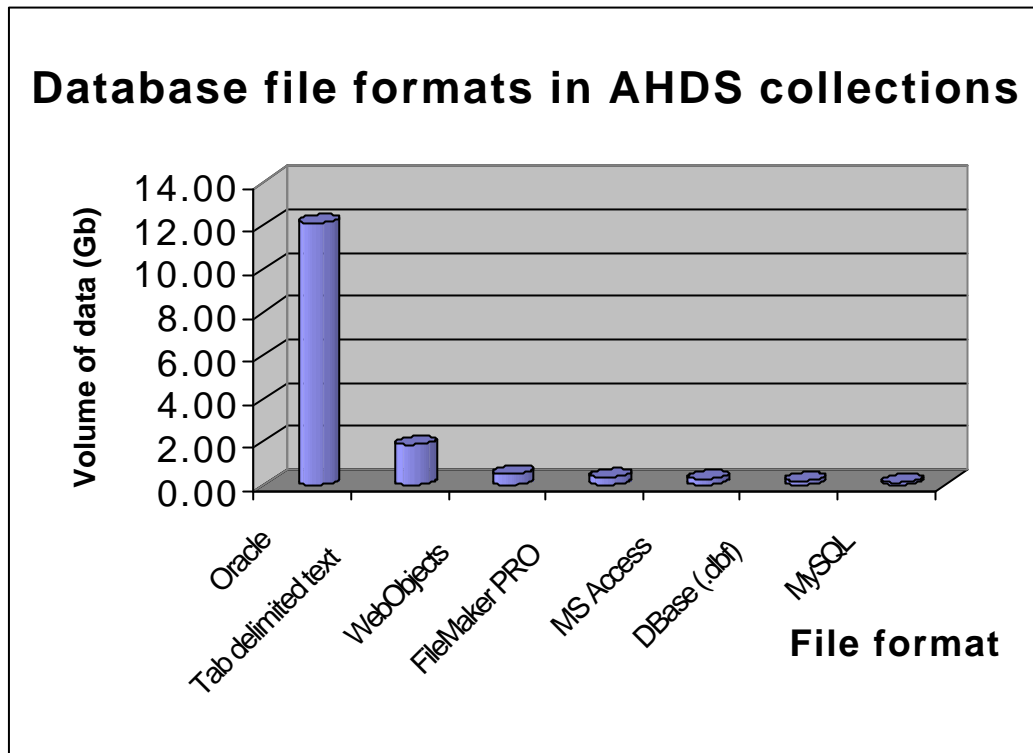


Figure 4. Database file formats in the AHDS collections.

A separate question in the file format survey asked about practices of describing the database structures and their other functionalities. The answers are listed in a separate section of the table above.

Only two database file formats were mentioned by the Service Providers as potentially problematic on the grounds that the Service Provider had not obtained a fully licensed copy of the database management system. Given that the same file format is also preserved by other Service Providers, maintaining access to this data should not be difficult as long as the software program exists within the AHDS.

As the data volume in Oracle database demonstrates (see Figure 4) how the AHDS Service Providers' collection composition can be considerably influenced by one study or deposit that can have serious implications on the preservation practices. Some of the image and video resources, as well as the 1881 census dataset, are further examples of such single acquisitions that can double the size of the collections.

4.2.4 Risks and factors affecting the preservation of database data file formats

The main format chosen for preservation of database data is a standard and simple delimited text format. This presumes that in addition to the data the structure and the interface of the database have been described and saved alongside the data. Based on the survey it cannot be said that the practice of describing database structure follows a consistent and uniform standard (one Service Provider responded that they do not store the database structure separately from data).

Other, non-standard formats chosen for deposit and preservation are generally well-compatible with other database systems and as a minimum the data from a database can be retained in these formats at very low level of risk. Functionalities of database management systems, like reports, data entry forms, indices, etc., cannot be preserved for long term with the current AHDS preservation strategies and should either be described in metadata, saved as snapshot images (e.g., in PDF), or both.

A large number of deposit formats for databases increases the time and costs associated with validation of datasets, or the evaluation process by which the Service Providers ensure that the datasets that were received correspond to the documentation provided by the depositor. Manual methods and automated validation tools have been developed by the AHDS Service Providers and other preservation services for processing database data and these can be shared by other archives.

4.2.5 Conclusions and recommendations

The main bulk of database data in AHDS holdings are saved in a low-risk, low-level simple format that is safe from the long-term preservation point of view. The chosen dissemination formats can potentially become labour-intensive for Service Providers when they have to be created from simple tab-delimited files that are used for preservation.

Recommendations

In the long term, the AHDS would benefit from a more unified and standardised description of database structures and functionalities. This would allow using partial automation when creating new dissemination formats and unify the procedures level data processing practices across the AHDS. Examples of such description are formalised SQL queries (the HDS is already using the SQL Create Tables queries for describing the structure and interrelationships of tables in a database) and the basic structure for a descriptive ‘metadatabase’ that describes and defines the structure of archived databases (developed by a Nordic TEAM-1 project²⁰).

Given that the majority of databases deposited with the AHDS derive from PC-based database management systems, it would be feasible to co-ordinate among the Service Providers (or through a centralised AHDS preservation facility) that at least one copy of the more popular PC database management software installation sets are retained (and the operating systems that they require). Such ‘technology preservation’ would lower further the risks associated with long-term preservation of database data.

The AHDS should also co-ordinate the establishment of a Technology Watch service that would monitor the developments in database technology, most notably in the object-oriented and web-based databases, and solutions for their successful preservation.

²⁰ see “To Preserve and Provide Access to Electronic Records”, 1996, p. 7

4.3 STATISTICAL AND SPREADSHEET DATA

4.3.1 A spreadsheet is a two-dimensional matrix of cells that typically contain numeric data. Each cell also contains a unique identifier (usually its X and Y co-ordinates) and may contain a formula to execute a mathematical function and formatting for the presentation of the cell content. A spreadsheet is self-referential in the sense that information about the content of each row and column is embedded in the worksheet, along with the actual data and any formula used for mathematical computations. Consequently, spreadsheets can be characterised as self-contained but dependent upon a specific software application to retrieve and display data.²¹

4.3.2 Presently, no economically realistic standardised way exists to maintain the calculation capability of spreadsheets for long term. As for databases, no high-level standardised formats are in use for spreadsheets and the safe preservation option consists of saving separately the data and the description of any functionalities offered by the creating software.

Microsoft Excel can be regarded as a *de facto* proprietary standard for spreadsheet files,²² but a number of software vendors provide compatibility of Excel files across different programs and their versions (backward compatibility). This means that for the short term it is possible to automatically convert a spreadsheet in an older version to a newer one with relative ease. The Data Interchange Format (DIF) — a generic low-level file format for data transfer — can be used to ensure the processibility of spreadsheets through preservation of the computational formulae and cell entries. However, a number of key spreadsheet features, such as comments about specific cell entries, hyperlinks and type font will be lost.²³

Recent versions of MS Excel include a feature of saving spreadsheets as XML files. Very little testing or research has been published as to the suitability of such XML spreadsheet files for long-term preservation, or whether the primary purpose of this format is to facilitate cross-platform interchange in the short term. The Microsoft version of XML coding is generally not regarded as an open XML that is safe for long-term preservation.

4.3.3 The AHDS Service Provider survey identified the following file formats as preferred and acceptable for statistical and spreadsheet data (see Table 6 below).

²¹ Ch. Dollar, "Authentic Electronic Records: Strategies for Long-Term Access", 1999, p. 163

²² DLM-Forum, "Guidelines on best practices for using electronic information", 1997, p. 42

²³ Ch. Dollar, "Authentic Electronic Records: Strategies for Long-Term Access", 1999, p. 163

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
Tab delimited text	0.01	Tab delimited text	0.49	Tab delimited text		0.50
Comma delimited text (.csv)	0.01	ASCII	0.29			0.30
MS Excel	0.101	MS Excel	0.11	MS Excel (v. 97 <)		0.112
Lotus 123	0.02		0.05			0.07
Quattro Pro						
SPSS portable (.por)		SPSS portable	0.18	SPSS portable		0.18
SPSS system (.sav)						
				SPSS export	0.02	0.02
Acceptable formats						
		Fixed width text	0.88			0.88
.tab						
.csv						
Quattro Pro						
STATA (.dta)						
SAS transport file					0.01	0.01
Expected new formats in the near future						
-						
Problematic formats for the Service Providers						
-						

Table 6. Statistical and spreadsheet data file formats in AHDS collections.

The table includes both spreadsheet file formats as well as those used for statistical data (e.g., SPSS, SAS). The fixed width file format is only used as a legacy format for statistical data at the HDS and is no longer created for preservation nor offered to users (except on request) and is not encouraged for deposit. The long-term preservation of these legacy files will not pose a problem as they are simple and ASCII-based, but the documentation that enables the interpretation of data must always be preserved alongside.

The AHDS Service Providers did not name any new or problematic file formats for this data type.

The overall distribution of file formats is as follows (see Figure 5 below):

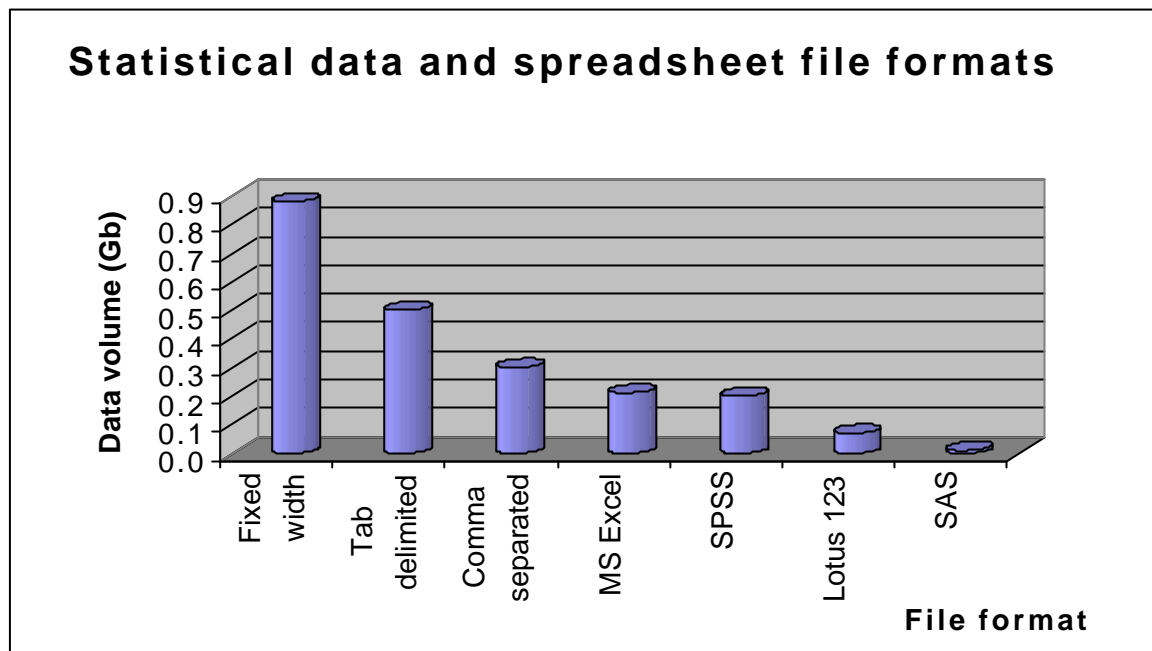


Figure 5. Distribution of statistical and spreadsheet data between file formats.

4.3.4 Risks and factors affecting the preservation of spreadsheet file formats

The primary file formats chosen for preservation of statistical and spreadsheet data are the safest from the long-term retention point of view: preservation of ASCII text based, delimited files is trivial. At the same time, it presumes a detailed level of documentation to accompany the data files that explain any calculations or other bases of cell contents in data. This makes the creation of such preservation formats labour-intensive to create and validate.

Comma delimited file format, or Comma Separated Value (CSV), where each cell entry is separated by a comma, has a high level of errors that can be introduced at the creation stage (e.g., comma is used to separate the decimal units or mark the thousands in a number). This risk can be lowered somewhat by including all cell entries in double quotes in a CSV file.

The long-term preservation of proprietary file formats (MS Excel, Lotus 123) presumes a migration strategy and migration path that is periodically reviewed to ensure timely conversions to newer versions of software or to new formats.

The relatively high level of compatibility between the proprietary spreadsheet formats and software programs that the AHDS has chosen, renders the creation of dissemination formats an easy and uncomplicated task.

4.3.5 Conclusions and recommendations

The file formats chosen by the AHDS to preserve the statistical and spreadsheet data are low-risk formats and can be considered safe, provided a migration strategy is in place and is being reviewed and followed for non-standard formats.

XML should be explored as a standardised format for preserving both data and descriptive elements of a spreadsheet together.

4.4 DIGITAL IMAGES

4.4.1 Digital image formats can be divided into bit-mapped graphics (also known as raster-based) and vector graphics formats. Vector graphics are mathematical representations of graphical elements that can render lines, colours, and shapes and are processible like ASCII text. Bitmap images are numerical representations of the variation of reflectance of a targeted area of picture elements (pixels) expressed as number of dots per inch (dpi). In the matrix of pixels each pixel contains the digital colour information sampled from its corresponding analogue original. The number of dots per inch measured horizontally and vertically determines the pixel size and each pixel can be represented by one to twenty-four bits (bi-tonal image, grey scale or colour image). The main advantages of bitmap image formats are that each pixel can be manipulated individually and that a high degree of photorealism is obtainable. Additionally, once the image has been loaded, displaying (as opposed to manipulating) the image is often much faster than with vector formats as it does not require very much processor power or re-calculations. The raster image file sizes tend to be large because of information about every individual pixel is stored in them. Because the storage requirements are considerable, compression techniques for reducing them have been developed.

The compression methods are of two varieties: lossless and lossy. Lossless compression uses algorithms that, typically, encode repeating elements (patterns) within an image (LZW, RLE, PKZip). For example, stretches of pixels that share the same colour are taken and stored in just two bytes: one for colour and the other for the number of adjacent pixels. Lossless compression means that after being compressed and then decompressed, an image is exactly the same as original. Compression ratios achieved are not very high, typically about 50-75% of the original file size. Lossy compression techniques are capable of much higher ratios, by removing some of the less useful information in an image (which is often not distinguishable to human eye). Lossy compression formats are, as a rule, not suitable for archival preservation because the image quality is reduced through compression.²⁴

4.4.2 Ensuring long-term access to bitmap images through technology generations poses several problems. One is that digital images typically carry no intelligence: a computer cannot “recognise” a portion of an image and then execute some operation or computation, which means that images always have to be supplanted with description and metadata. The more technical attributes of an image (e.g., height and width, bit depth, byte order, scanning resolution, etc.) can sometimes be included in an Image File Header (IFH).

Most image file formats are proprietary products that are supplied as part of an integrated digital imaging system. Such proprietary software requirement is not viable for a long term preservation strategy, particularly since given relatively stiff commercial competition the formats for image files are changing and will continue to change rapidly.

Currently, no internationally accepted standard exists for preservation of image file formats. The Tagged Image File Format (TIFF) is widely in use and is effectively a *de facto* standard, albeit being proprietary. The format was developed by Aldus and Microsoft to provide a basis for importing scanned images into desktop publishing packages. It is a lossless format but different versions offer a limited level of compression options (LZW/CCITT compression,

²⁴ UNESCO Memory of the World Programme, “Recommendations of the Committee on Technology for Consideration by the International Advisory Committee”, 1995, p. 9

JPEG compression was introduced in version 6). TIFF files can also contain contextual metadata (e.g., author information, copyright, etc.).²⁵

Joint Photographic Experts Group (JPEG) format is a lossy compression format and contained a proprietary (IBM) component which rendered it unsuitable for authentic long-term preservation of image data. The JPEG 2000 initiative was set up in 1998 to improve on the JPEG format by using better compression algorithms. The JPEG 2000 became an international standard ISO/IEC 15444-1:2000 (“JPEG 2000 image coding system”, with corrections and amendment enforced in 2002). Only the first part of the standard of the intended eight have been published so far.²⁶ The new format “desires” that both lossless and lossy compression be available at time of saving. The JPEG 2000 uses compression techniques based on wavelet technology and will produce images without the blockiness of the former JPEG format. It is also claimed that the new compression algorithm is 20% more efficient than the currently used one.

The development of the Portable Network Graphics (PNG) format was promoted by the World Wide Web Consortium²⁷ and it uses a lossless compression algorithm (LZW). It was intended to replace the proprietary Graphics Interchange Format (GIF) as an Internet image format, but it has not been universally accepted yet. The current official specification for the PNG format supports 48-bit colour and 16-bit greyscale encoding. However, current Internet connection bandwidths so limit the use of files containing so much information that only 8 and 24-bit versions are presently used for web graphics. PNG offers several features not available in GIF images and is independent of hard- and software platform. PNG can also contain searchable metadata in the form of keywords and text strings.

BMP is the MS Windows native bitmap image format. Although it supports run-length encoding (RLE), it is commonly used as an uncompressed format and file sizes can be quite large. Therefore it is seldom the choice for large high-resolution images, but it is supported by many image processing software packages.

Graphics Interchange format (GIF) was a popular and widespread image format, owned and used by CompuServe. It provided for LZW compression, interlacing, transparency and multiple images per file. Although being outphased by the PNG, many software tools still support GIF format and several shareware viewers / converters are available.

Vector image formats that are processible as ASCII encoded data and compared to bitmap images require very little storage space, nevertheless have some issues to note for long-term preservation. These will be reviewed under CAD data below in section 4.7.

²⁵ TIFF 6.0 Specification is available from <http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf>

²⁶ for more information see <http://www.jpeg.org>

²⁷ for more information see <http://www.w3.org/Graphics/PNG/>

4.4.3 The image file formats in AHDS collections are listed in Table 7 below.

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
TIFF	450	TIFF	28.51	TIFF		478.51
		TIFF (v. 6 <)	0.5			0.5
PNG	0.01	PNG	0.17	PNG		0.18
PDF	0.75	PDF	1.5	PDF		2.25
JPEG	0.25			JPEG	51.2	51.45
BMP	0.28					0.28
BIL						
CGM						
PCX						
PhotoCD						
Photoshop						
				PostScript		
				GIF		
DXF		DXF				
SVG		SVG				
Acceptable formats						
JPEG	0.07	JPEG		JPEG		0.07
GIF	0.32			GIF	1.9	2.22
Adobe Illustrator	4.96			Adobe Illustrator		4.96
BMP						
Photoshop						
ESRI shape files						
GeoTiff						
Expected new formats in the near future						
-						
Problematic formats for the Service Providers						
SVG						

Table 7. The image data file formats in the AHDS collection.

As the table demonstrates, TIFF is the file format with by far the largest volume of data — it forms ca 78% of the total AHDS collection. The HDS expects further 250Gb of TIFF data over the next year or two, by which time collections of other Service Providers will have grown, too and it is possible that in two years time the AHDS will hold 1 Tb of data in TIFF format. Handling such large data volumes and duplicating the provision of adequate level of preservation maintenance and back-up copies can become burdensome as well as resource-demanding for individual Service Providers — a centralised AHDS preservation facility equipped with appropriate technology is likely to offer savings in terms of labour, time and cost.

Several Service Providers noted that for preservation, the uncompressed TIFF and JPEG formats are used. JPEG is used as a preservation format only when the original data was deposited as JPEG files, the main function of this file format remains dissemination and its volume is projected to rise when VADS will adopt the policy of creating four derived JPEG images from each TIFF and the PADS website is redesigned.

The file format Photoshop in the table indicates vector image formats that the Adobe Photoshop software can read and import (OTA currently holds no image data but is prepared to accept all formats that the Adobe Photoshop can import).

The Scalable Vector Graphics (SVG) and DXF are vector graphics formats but no image data in these formats is currently in the AHDS collections, however, it is expected to appear in the future and two Service Providers listed it as a format that may become problematic. The preservation of vector graphics formats is discussed further under CAD data below.

The distribution of image file formats in AHDS collection according to data volume is presented as a table below:

File format	Data volume (Gb)
TIFF	479.01
JPEG	51.52
Adobe Illustrator	4.96
PDF	2.25
GIF	2.22
BMP	0.28
PNG	0.18

Table 8. Image file data volumes.

4.4.4 Risks and factors affecting the preservation of image data file formats

The only recommended format for archival preservation of image files is currently TIFF. Although it is a proprietary format and new versions of it are being developed and released, its specification has been published and it can be considered a semi-open standardised file format with low risk level for long-term preservation. In the absence of an effective method for retaining image files in a software independent format, TIFF offers probably the safest lossless preservation of image data.

The AHDS Service Providers have firmly chosen TIFF as the main preservation, although a small number of other formats are retained alongside. The file format questionnaire did not ask specifically about the practices of creating TIFF files for preservation, therefore, it is not possible to say how much of the TIFF data currently preserved was deposited and what is the proportion of other image file formats that are converted into TIFF. The latter is probably negligible.

JPEG is not considered an adequate long-term preservation format primarily because it is a lossy compression image file format, even though a version of it has become an international standard. It is, nevertheless, widely used and can be read and processed by most image processing software, which reduces the risk of losing access to the JPEG files in the long term. As an additional safety measure, conversion of JPEG and GIF files to PNG and/or SPIFF formats could be considered.

4.3.5 Conclusions and recommendations

The AHDS has chosen TIFF as the main preservation format for image data and a small number of other, less safe, but well-compatible formats, which yields the overall preservation strategy to be a low-risk one. However, the volume of image data when using uncompressed formats may potentially become burdensome to handle.

Recommendation:

A centralised AHDS preservation facility should be considered for handling the preservation of large volume of image data, especially since it is all using the same file format.

4.5 DIGITAL VIDEO

4.5.1 In essence, digital video is comprised of individual images that are sequenced and played at a given speed (digital video is often referred to as ‘moving image data’). Video file formats are storage-intensive because of the amount of information stored for each image, even with using compression techniques. Digital video and audio use codec algorithms to minimise the amount of storage space required for a file. The compression works by eliminating redundancies in data that are decompressed for viewing. Video codecs typically work by discarding colour information that human eye is not sensitive enough to detect and by applying various temporal compression algorithms whereby the only information recorded is the information that changes from one frame to the next. The amount of colour a particular codec stores is referred to as its colour-sampling ratio.

Numerous strategies exist to reduce the number of bits required for digital video, from relaxed resolution requirements to lossy compression in which some information is sacrificed in order to reduce significantly the number of bits used to encode the video. Motion Picture Experts Group-1 (MPEG-1) and MPEG-2 are two such lossy compression formats.

4.5.2 The Motion Pictures Expert Group (MPEG) format enables the compression of bit streams containing moving images and audio using JPEG compression on individual frames of moving images and other lossy compression techniques to compress data between frames. Numerical suffixes of MPEG formats indicate the efforts of the Moving Pictures Expert Group that are compartmentalised into different standards targeting different markets. The first MPEG was developed with the stated goal of creating a digital video compression format that would rival “VHS quality” at a bitrate of 150 Kilobytes per second (1.2 Mbits/sec). That matched the bandwidth limits of then-current 1x CD-ROM drives and VideoCD players, and provided video resolution of 352 x 240 at 30 frames per second (fps). The resulting format MPEG-1 (ISO/IEC 11172:1993, in 5 parts) is using heavy compression techniques that are the foundation for both MPEG-2 and MPEG-4 as well. The difference is that those later formats target different bandwidths and different needs. With MPEG-2 (ISO/IEC 13818:1995, in 9 parts), the MPEG ISO sub-committee strove to achieve the best quality video for broadcast television without the narrow bandwidth parameters established for MPEG-1. (The name MPEG-2 makes sense chronologically, but it is not a replacement for MPEG-1.) MPEG-2 video quality far exceeds that of MPEG-1, but so does its demand for bandwidth. It offers resolutions 720 x 480 and 1280 x 720 at 60 fps, with full CD quality audio.

The MPEG-4 (ISO/IEC 14496:1999 “Very-low bitrate audio-visual coding”, in 10 parts) standard targets the more restrictive bandwidths of Internet and wireless networks. MPEG-4 quality can be quite similar to MPEG-1, but it supports reduced frame rates (dropping from 30 fps to 15 fps can save half the required bandwidth). MPEG-4 is the only streaming digital video format that is an international standard and not the property of one company, like are for example the RealVideo, Windows Media, and QuickTime with Sorenson compression.

Two further MPEG standards are in the proposal stage: MPEG-7 that centres around describing media material contained in the file for archiving and retrieval purposes, while MPEG-21 focuses on digital media compatibility.²⁸ MPEG-7 is formally called ‘Multimedia Content Description Interface’, a means of attaching metadata to multimedia content. MPEG-7 does not define a monolithic system for content description but rather a set of methods and

²⁸ for more information see: <http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm> and <http://mpeg.telecomitalia.com/standards/mpeg-21/mpeg-21.htm>

tools for the different viewpoints of the description of audio-visual content. It uses XML Schema as the language of choice for the textual representation of content description and for allowing extensibility of description tools.

Apple's moving image format QuickTime is the oldest of the video file formats and is used to store QuickTime movies, other audio, video and animations, as well as other time-based data. QuickTime format itself does no video compression, though it supports several compression formats. The current default compression type is the Sorenson Video codec. A QuickTime file always contains metadata, a collection of headers which describe the kind of media present in the movie (audio, video, animation, etc.), the precise data format of each chunk of media data, the position in the movie timeline of each chunk of media data, and an index of the physical location (file and file offset) of every chunk of media data. A QuickTime file stores the description of the media separately from the media data. The media data is all of the actual sample data, such as video frames and audio samples. The media data may be stored in the same file as the QuickTime movie, in a separate file, or in several files.²⁹

Because the file format can be used to describe almost any media structure, Apple suggests it as an ideal format for the exchange of digital multimedia between applications, regardless of the platform on which the application may be running.

The QuickTime file format has been used in the development of the MPEG-4 standard.

4.5.3 The file format survey identified the following file formats as preferred and acceptable for depositing, storing and disseminating the digital video resources:

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
MPEG		MPEG		MPEG		
MPEG-1						
MPEG-2		MPEG-2		MPEG-2		
MPEG-4		MPEG-4				
Quicktime	0.56			Quicktime	0.62	1.18
Real Video				Real Video		
		SMIL				
Acceptable formats						
Quicktime		Quicktime		Quicktime		
AVI		AVI		AVI		
Expected new formats in the near future						
Quicktime						
Problematic formats for the Service Providers						
Quicktime						

Table 9. Digital video file formats in AHDS collection.

Thus far, the AHDS Service Providers have collected relatively little digital video data and it is all in only one file format — Quicktime. More digital video resources in the same file format are expected to be acquired in the near future.

Digital video files are generally large in size and storage space questions have been commented on by the Service Providers in the file format survey. A useful role for a

²⁹ for more information see: <http://developer.apple.com/techpubs/quicktime/qtdevdocs/OTFF/qtff.html>

centralised AHDS preservation facility was suggested to be precisely to deal with large data volumes and potentially problematic file formats.

4.5.4 Risks and factors affecting the preservation of digital video file formats

MPEG is currently the only non-proprietary compression algorithm available for digital moving images, leaving practically no other choice for a preservation format. At the same time, MPEG's position among popular formats that are used for delivery of video over networks is low and it is likely that the AHDS Service Providers will be offered video data in these formats and not in MPEG formats. The low- to medium risk preservation strategy for video data in these formats (e.g., RealVideo, Windows Media, etc.) is to preserve the software necessary for rendering the files alongside the archived video data resources. This will add to cost and complexity of the preservation process, but will ensure the continuing usability of collections. Given the relatively good backward compatibility offered by the digital video playback software, it is likely that it will not be necessary to preserve every version of each software program. Nevertheless, the file format versions that the preserved software version can read in and output must be recorded in the preservation metadata (cf. "AHDS Preservation Metadata Framework", element no. 12).

All digital video data in the AHDS collection is currently stored as Quicktime files, even though this file format has not been listed among the preferred preservation formats for digital video. Conversion between digital video file formats remains a problematic topic and has not been sufficiently tested for preservation purposes.

4.5.5 Conclusions and recommendations

The AHDS digital video collection is currently stored in a (published) proprietary format that the data was deposited in and no conversion or migration strategy exists for long-term preservation of the format. The high risk of losing access to the data resources in the future associated with proprietary file formats can be reduced, up to an extent, when originating software is archived together with the data files. The AHDS must establish this practice as the minimum preservation strategy for its digital video resources.

Recommendations

AHDS should also investigate the conversion possibilities between different formats (particularly conversion into MPEG-2).

MPEG-4 format should only be used as a preservation format when video data is deposited as MPEG-4 files.

4.6 DIGITAL AUDIO

4.6.1 Digital audio data are binary representations of the analogue sound signals. The conversion from analogue to digital – sampling – is achieved by measuring the level of sound on the wave and assigning a numeric value to it. The frequency of sampling (sampling rate) determines the quality (e.g., less choppiness) of the digital sound and sample resolution refers to the number of bits recorded per sample. Similar to bitmap images, the higher the number of bits per sample, the greater the capability to record and reproduce the full dynamic range (loud to soft) of the audio. Digital samples are discreet entities but the resulting sound will be smooth if the individual snapshots are taken at sufficient frequency: for voice communication, a sampling rate of 8,000 per second is considered adequate; for non-voice audio, the compact disc sampling rate of 44,1 kHz is considered as high and 11,25 kHz as low³⁰ (although some user communities require much higher sampling rate than 44,000 – suggested frequencies go up to 192,000).

For most audio signals there is considerable redundancy that is usually removed with a compression algorithm. Compression can be lossy or lossless, the latter resulting in large sound files.

4.6.2 Digital audio file formats can be categorised into two broad categories: self-describing formats, where the playing device parameters and encoding are made explicit in some form in the file header, and raw formats, where the device parameters and encoding are fixed.

The Waveform Audio File Format (WAVE, .wav) is a proprietary standard developed by Microsoft and IBM as part of the Resource Interchange File Format (RIFF) for MS Windows 3.1.³¹ WAVE support was built into MS Windows 95 and it has become a widely used format since. A variety of applications now support .wav, as do additional operating system platforms such as Apple Macintosh.

WAVE encoded audio data is preceded in the file by a Format Chunk which identifies the required WAVE format, the number of channels, the sampling rate, the type of block alignment. The default WAVE format is the Microsoft Pulse Code Modulation (PCM) format; alternative formats identify the use of A-law and U-Law PCM variants, and Adaptive Differential Pulse Code Modulation (ADPCM). The Data Chunk of a WAVE file can include Fact Chunks containing additional facts about the data, Cue-Points Chunks identifying synchronisation points, Playlist Chunks defining the playing order, and Associated Data Chunks containing labels, notes and related text (which can be language dependent). .wav has become a common interchange format for audio files transmitted over the Internet.

The AU audio file format was developed by Sun Microsystems/NeXT for Unix platform. It is a popular format for sound and audio samples. The format is not highly compressed and it encodes audio data in three parts: a header (containing fields that describe the audio encoding format); a variable-length information field (in which, for instance, ASCII annotation may be stored); and, the actual encoded audio. The recent versions of Internet browser software packages, such as Netscape Navigator, are able to play AU files, which have the extension .au.

The MPEG standards (see also previous section on digital video) can also be used for encoding and storing digital audio. MPEG-1 (ISO/IEC 11172-3:1993) audio signals can be encoded in single channel, dual channel (two independent signals), stereo or joint stereo

³⁰ Ch. Dollar, "Authentic Electronic Records: Strategies for Long-Term Access", 1999, p. 157

³¹ for more information on the format see: <http://www.techweb.rfa.org/docs/RIFFWAVE.htm>

formats using pulse coded modulation (PCM) signals sampled at 32, 44.1 or 48 kHz. MPEG-2 is a backwards-compatible multi-channel extension of the MPEG-1 audio standard – it provides a greater range of sampling formats. The MPEG Audio Layer standard is better known as MP3 format which is widely used for the interchange of music over the Internet. (There is no MPEG 3 standard as one of the AHDS Service Providers had used in the file format survey responses, MP3 refers to Audio Layer 3 in the MPEG standard.) MP3 files are tightly compressed but preserve the original quality of the recording.³²

The Real Audio file format (.ra, .ram) is proprietary: the software includes multiple codecs, all of which are proprietary to Real. Real Audio is widely used for streaming audio signals over the Internet. It provides (depending on the version) from AM-quality over 14.4 Kbps up to near-CD quality over ISDN and LAN connections.

Ogg Vorbis is a new (and free) digital audio codec and format. It is trying to rival MP3 by offering mid to high quality audio encoding for Internet distribution. Ogg Vorbis is an open, non-proprietary, patent-and-royalty-free, general-purpose compressed audio format. The format specification was published in July 2002. Xiph.org's Vorbis software libraries are distributed under a BSD-like license.³³

4.6.3 The AHDS collections currently hold and are prepared to accept the following digital audio formats:

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
MP3		MP3		MP3		
WAV	0.1	WAV		WAV		0.1
MPEG		MPEG		MPEG		
AU		AU				
Real Audio	0.075			Real Audio		0.075
		Ogg Vorbis		Ogg Vorbis		
Quicktime						
Acceptable formats						
Real Audio		Real Audio		Real Audio		
				MP3		
Expected new formats in the near future						
Ogg Vorbis						
MP3						
Problematic formats for the Service Providers						
-						

Table 10. Digital audio formats in AHDS collections.

There are very few audio data resources currently deposited with the AHDS Service Providers, but both HDS and OTA are expecting more acquisitions in the near future. The likely formats for future deposits include Ogg Vorbis and MP3.

³² for more information see: <http://www.iis.fraunhofer.de/amm/techinf/layer3/index.html>

³³ for more information see: <http://www.xiph.org/ogg/vorbis/docs.html>

4.6.4 Risks and factors affecting the preservation of digital audio file formats

The open and standardised Ogg Vorbis and MPEG-based formats will pose less problem for long-term preservation than the proprietary formats. Both, however, are prone to rapid change and development since digital audio is still a competitive area of software market. The best strategies for ensuring continuing access to such formats is the same as discussed for digital video data – to preserve the originating software program alongside the data resource.

Since improving access to audio resources is one of the main driving forces for their digitisation, it is likely that the future deposits of audio data with the AHDS Service Providers are going to be in formats that are suitable for distribution and streaming over the Internet.

4.6.5 Conclusion and recommendations

Thus far, the AHDS only holds proprietary digital audio formats and in order to lower the risks associated with preservation of such formats, the original software should be archived together with the audio data (cf. recommendations for digital video).

AHDS should establish a Technology Watch service that would focus on developments in the digital video and audio formats market and make recommendations for necessary documentation (metadata) that should be requested from the depositors at the ingest stage and archived with the deposited data resources.

4.7 CAD

4.7.1 Computer-Aided Design (CAD) software is widely used by different communities, including scholarly researchers to design or document physical structures and objects. CAD programs have developed from creating two-dimensional drawings to allowing three dimensional designs and visualisations of complex objects. CAD applications describe the drawn models in layers that can be displayed separately or together and generally store the graphical information about each layer as vector graphics: lines, arcs and circles are stored using mathematical formulae (vectors) so that they can be represented at any scale.

In vector graphics applications, lines and shapes are associated with information that specifies size, shape and position relative to the overall image, colour and other attributes. Vector graphics require relatively little storage space and are often processible as ASCII encoded data.

The geometric objects represented in a CAD system can be described in detail by data in external tables that some applications allow to be attached to the CAD model. Modern CAD programs also hold the co-ordinates and measurements for the data-points used to create models in separate data tables. CAD files that constitute a complex object or a project can be linked and cross-referenced (xref files) to make up a composite model.

4.7.2 Popular CAD software uses proprietary file formats and these do not necessarily transfer successfully between different programs. At present the most commonly used CAD software is AutoCAD, developed by Autodesk.

The AutoCAD's native format for CAD models DWG is perhaps the most widely used CAD file format and has become a *de facto* industry standard. It is the proprietary format of AutoCAD, but other software manufacturers have made their products compatible for reading and using the .dwg files. There are attempts to make DWG a public standard.³⁴

Drawing Exchange Format (DXF) is an output format used for exchanging (Auto)CAD data. It, too, is a proprietary but open standard maintained and developed by Autodesk. It has changed slightly with virtually every new release of AutoCAD software. DXF is used for interchange between CAD and other vector graphics packages, particularly on PC and UNIX computers. Most PC drawing and illustration software supports the import and export of this ASCII-based format. The main disadvantage that has been commented on with this format is the slow loading and saving speed thanks to the large volume of co-ordinate data in ASCII form that needs to be read or written. There are also incompatibility problems with .dxf in that software packages which do not support particular versions of .dxf may still import the data but incompletely.

The Drawing Web Format (DWF) format is generally used to publish CAD models on the Internet. The .dwf is a highly compressed file that is created from a .dwg file. It is not recommended that CAD files are either stored or archived as .dwf.³⁵

The Scalable Vector Graphics (SVG) format is based on XML and it is highly compact. It provides a language for describing instructions for two-dimensional graphics in XML and is thus ASCII-based. SVG allows for three types of graphic objects: vector graphic shapes (e.g.,

³⁴ see: <http://www.opendwg.org/downloads/guest.htm#dwgspec>

³⁵ <http://ads.ahds.ac.uk/project/goodguides/cad/sect45.html>

paths consisting of straight lines and curves), images and text. Graphical objects can be grouped, styled, transformed and composited into previously rendered objects. SVG drawings can be dynamic and interactive; they are compatible with HTML, CSS, JavaScript and CGI. The Document Object Model (DOM) for SVG allows for straightforward and efficient vector graphics animation via scripting.

4.7.3 The only AHDS Service Provider that has accessioned CAD data is ADS:

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
DXF	0.393	DXF		DXF		0.393
DWG	0.248	DWG		DWG		0.248
				DWF	0.144	0.144
Acceptable formats						
native format				native format		
Expected new formats in the near future						
-						
Problematic formats for the Service Providers						
-						

Table 11. CAD data formats in ADS collections.

All CAD data in AHDS collections is held by one Service Provider and is in formats that are produced by one software program. The formats can be considered to be *de facto* industry standards.

4.7.4 Risks and factors affecting the preservation of CAD data file formats

Given that there is no standard format for exchanging CAD data between different software packages that could be considered completely safe for preservation purposes, the best strategy on offer is to use the most common file formats for storing the CAD data. The ADS has followed this advice.

The ADS own guidance points at incompatibility problems between different file formats and even between versions of the same file format from the same manufacturer and recommends that CAD files are to be saved in the latest possible version of DWG and DXF together with adequate documentation.³⁶ Hence, continuous migration of CAD files to new versions of the AutoCad's proprietary formats will be necessary.

A consortium in Japan has developed a long-term archiving standard for CAD using the ISO 10303 standard description (STEP) for describing CAD files.³⁷ The archiving standard has emerged from the manufacturing and engineering industries where there is a big demand for a solution to long-term CAD data archiving that would be independent from the specific CAD systems. In particular, the standard is using these parts of the international STEP standard:

³⁶ <http://ads.ahds.ac.uk/project/goodguides/cad/sect51.html>;
<http://ads.ahds.ac.uk/project/goodguides/cad/sect45.html>

³⁷ see: <http://www.mosla.org>

- ISO 10303-203 “Industrial automation systems and integration - Product data representation and exchange - Part 203: Application protocol: Configuration controlled design”, 1994 (called STEP/AP203)
- ISO 10303-202 “Industrial automation systems and integration - Product data representation and exchange - Part 202: Application protocol: Associative draughting”, 1996 (called STEP/AP202)
- ISO 10303-21 “Industrial automation systems and integration - Product data representation and exchange - Part 21: Implementation methods: Clear text encoding of the exchange structure”, 1994, 1996 (called STEP/Part21).

Research for this consultancy report did not reveal any software packages that are supporting this standard.

4.7.5 Conclusion and recommendations

The ADS is well aware of the problems associated with the CAD file formats and their preservation needs. So long as the number of CAD file formats in AHDS collections remains small, the problem with formats being proprietary, updated irregularly and not open can be overcome by careful research into file format migration paths and regular conversion to newer formats. However, when the data volume and/or number of formats grows, a standardised format or description will have to be found.

4.8 GIS

4.8.1 A Geographical Information System (GIS), or a Spatial Information System, is a computer system for capturing, storing, checking, integrating, manipulating, analysing and displaying data related to positions on the Earth's surface. Typically, it is used for handling maps of one kind or another.³⁸ GIS can incorporate a series of different maps (vector and raster) linked to data and overlying one another in layers or coverages. The graphic object (feature on a map) and data can be taken together as parts of a set on which the GIS can perform mathematical and other functions. In a GIS the connection between spatial information and tabular data is more robust and more central to the functions than in a CAD system. However, GIS cannot usually be used to model complex three-dimensional objects adequately.

There are two major methods of storing mapped information that differ in how they conceptualise, store and represent the spatial locations of objects. Geographic Information Systems which store map features in vector format store points, lines and polygons. Raster Geographic Information Systems, which store map features in raster or grid format, generalise the location of features to a regular matrix of cells.

Some GIS packages are better at (or are only for) processing raster graphics data (e.g., GRASS, IDRISI), whereas some other are better at vector data (e.g., Arc/Info on a PC) but cannot manipulate raster data. Many Geographic Information Systems handle both vector and raster data (e.g., ArcView, Intergraph) from a wide variety of sources including satellite imagery, cadastral information, hand digitised maps and scanned images.

4.8.2 GIS software was first developed for and used on UNIX computers that offered more processing power than a PC computer. The PC versions of GIS appeared somewhat later and usually offered a more limited functionality. GIS data formats contain complex, multi-theme collections of spatial information that is not just sufficient to draw maps, but also contain necessary ancillary data about the features included (in space and time). The GIS file formats have remained proprietary to the software that generated them and so far, no consensus on standardised formats for preserving GIS data for long term exists.

ESRI's Arc/Info is one of the more widely used GIS packages. The UNIX and Windows NT releases provide comprehensive raster and vector processing capabilities, but the PC release is vector only. The ArcInfo's native file formats can be read into some other GIS packages, but the ArcInfo export format (.E00) is more commonly used for it. The E00 is an exchange and publication format that, however, must be converted into shape file (SHP) before it can be used by some software.

Although GIS software vendors continue to push other, more elaborate formats, the ArcInfo shape file appears to be the current *de facto* standard. It is not, however, easy to create, so only a few software programs have SHP export capability. When used in ESRI ArcInfo, ArcView and ArcExplorer programs, the SHP file is accompanied by the following auxiliary files:

- Shapefile database files (DBF)
- Shapefile index files (SHX)
- and sometimes by auxiliary files that store the spatial index of the features (SBN & SBX).

³⁸ <http://ads.ahds.ac.uk/project/goodguides/gis/sect72.html>

When used in ERDAS Mapsheets Express, two additional header files are usually created:

- STY - Formatting and georeferencing information
- LAB - Labelling information

The Shapefile spatial data format is open and published by ESRI.³⁹

Geographic Resources Analysis Support System (GRASS) and MOSS are public-domain raster GIS software developed by different U.S. government agencies. IDRISI is a raster-based commercial GIS package with its own file format.

MapInfo is a GIS program for PC computers that can produce geographical data for export in MapInfo Interchange Format MIF/MID. MapInfo exports data in two files – the graphics reside in a .MIF file and textual data is contained in a .MID file. The textual data is ASCII-based delimited data, with one row per record; it is editable, relatively easy to generate, and works on all platforms supported by MapInfo. The MID file is an optional file. The MIF file has two areas – the file header area and the data section. Information on how to create MapInfo tables is contained in the header; the graphical object definitions are in the data section.⁴⁰

Neutral Transfer File Format (NTFF) is an implementation of British Standard BS 7567,⁴¹ used for the transfer of geographic data. It provides a means for describing the contents of data records but does not define their contents. NTF provides:

- media-independent file and data record descriptions for the exchange of geographical data;
- a description of data elements, vectors and arrays containing character strings and numeric forms, and the relationship between data elements;
- volume and header information to enable data interchange to occur with minimal specific external description;
- five levels of interchange, the use of which depends on the complexity of the interchanged data.

Spatial Data Transfer Standard (SDTS) was developed to allow U.S. Federal agencies to share spatial data among applications which use different hardware, software, and operating systems. It is now an U.S. standard⁴² that specifies a structure and content for spatial data designed to support the transfer of different types of geographic and cartographic data. The standard includes seven parts that cover both raster as well as vector data.

The Vector Product Format (VPF) is a standard format⁴³ designed to be compatible with a wide variety of applications and products. VPF uses tables and indexes that permit direct access by spatial location and thematic content and is designed to be used with any digital geographic data in vector format that can be represented using nodes, edges, and faces. VPF defines the format of data objects, and the geo-relational data model provides a data organisation within which software can manipulate the VPF data objects.

³⁹ <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>

⁴⁰ for the MIF/MID format specification see Appendix J in:

http://www.mapinfo.com/free/docs/mipro/mipro_70_users.pdf

⁴¹ see: <http://bsonline.techindex.co.uk>

⁴² see: <http://mcmweb.er.usgs.gov/sdts/standard.html>

⁴³ see: <http://164.214.2.59/publications/specs/printed/VPF/vpf.html>

4.8.3 The only AHDS Service Provider that currently holds GIS data is the ADS. Their list of file formats includes the following:

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Preferred formats						
ArcInfo export	0.268	ArcInfo		ArcInfo		0.268
DXF		DXF		DXF		
DWG		DWG		DWG		
ArcInfo Shapefile	0.01					0.01
ArcInfo ungen						
ArcView (< v. 3)						
Idrisi (< v. 3)						
MIF/MID						
NTFF						
SDTF						
MOSS						
VPF						
Acceptable formats						
-						
Expected new formats in the near future						
-						
Problematic formats for the Service Providers						
-						

The ADS is holding GIS data only in ArcInfo formats which can be considered *de facto* standards.

4.8.4 Risks and factors affecting the preservation of GIS data file formats

In the absence of an openly defined standard format that would be supported by most software packages, the safest available strategy for long-term retention of GIS data is to use several *de facto* standard formats and migrate to new versions of formats and/or software as the need arises. The ADS has followed this strategy which, however, will be an expensive preservation practice over the long term.

4.8.5 Conclusion and recommendations

The ADS has issued best practice guidance on GIS data and their archiving. It is following its own advice and the small number of GIS datasets that the ADS currently holds can be preserved usable with the current preservation method. The basic text nature of the format makes it less complicated to preserve and process. Possibilities for conversion to other, standard formats should be explored for the future.

4.9 OTHER FILE FORMATS

The file format survey identified a number of other data types and file formats in the collections of the AHDS Service Providers. Since at present there is no or very little data deposited in these formats with the Service Providers and since only a few of these formats are open standards, they are discussed here collectively.

4.9.1 The data types and file formats covered in this section include:

Deposit formats	Volume	Preservation formats	Volume	Dissemination formats	Volume	Total
Multimedia and virtual reality						
SMIL		SMIL		SMIL		
VRML 2.0		VRML 2.0		VRML 2.0		
QTVR						
Macromedia Director						
Asymetrix ToolBook						
Geophysics						
AGF		AGF		AGF		
Contours	0.62					0.62
Geoplot	0.011					0.011
Other						
Web sites						
Executables	0.025					0.025

Table 12. Other data types and file formats that AHDS Service Providers accept.

The ADS holds some Geophysics data which is using software and formats specific to archaeology applications. The VADS has one executable computer assisted learning package in their collections. The other formats are listed as potential only for future acquisitions.

The formats and data types, and preservation issues related to them, will be looked at individually in the next four sections.

4.9.2 Preservation issues related to multimedia and virtual reality data are similar to other complex data types where proprietary software and file formats predominate and the software market is still developing and competitive. Archiving the virtual reality data means archiving its constituent parts: the files that make up the virtual ‘world’ (images, CAD files, etc.) and the data files associated with these, as well as the supporting documentation. Each of these virtual reality elements may have to be converted into standardised format for preservation, supplementing the preservation version with detailed description of how to put the elements back together to re-create the application.

In 1998, the W3C Working Group on Synchronised Multimedia (SYMM) published the Synchronised Multimedia Integration Language (SMIL) 1.0 Specification. SMIL defines a set of XML elements that can be used to group media object elements for parallel or sequential display. It also defines a set of media object attributes that can be used in conjunction with switch and link elements to control the presentation of media objects. Facilities are provided for controlling the duration of an object’s display, for repeating a synchronised set of media

objects and for identifying the region of the rendering surface to be assigned to each media object.

The timing extensions proposed in the HTML+TIME specification formed the basis of the Timing and Synchronisation module of the SMIL 2.0 Specification, which is designed to allow reuse of the SMIL syntax and semantics in XML-based languages, in particular those that need to represent timing and synchronisation. The specification is defined in terms of the following modules:

- Animation Module
- Content Control Module
- Layout Module
- Linking Module
- Media Object Module
- Metainformation Module
- Structure Module
- Timing and Synchronisation Module
- Time Manipulations Module
- Transition Effects Module

The specification also defines a SMIL Document Object Model that can be used to manipulate SMIL information sets.⁴⁴

The Virtual Reality Modelling Language (VRML) has become a *de facto* standard for 3D objects in the web environment. VRML is built on the Open Inventor technology from Silicon Graphics (the first VRML specification was defined as a subset of Silicon Graphic's Inventor File Format). It provides a graphical mark-up allowing links to other VRML and HTML resources and a language allowing users to interact with computer-generated 3D images. Interaction is defined in terms of translation and rotation of 3D images, control of lighting and perspective, and changing of textures and other display properties under the control of application-specific user interfaces. The standard VRML 97 specification introduced the concept of timed events being used to create routes within a 3D view. The standard has been published as ISO/IEC 14772:1998 'Information Technology – Computer Graphics and Image Processing – Virtual Reality Modelling Language (VRML)'.⁴⁵ Some public domain viewers for VRML files are available.

In July 2002 a final working draft for an extension to VRML 97, the Extensible 3D (X3D) standard, was published by the Web3D Consortium. As well as permitting time-based discrete and continuous changes to be defined, X3D allows event processing based on route semantics (event paths), loops and fan-in/fan-out events that depend on or generate multiple events. A TimeSensor allows looped cycles to be managed.⁴⁶

Apple's QuickTime VR (QTVR), is actually not a 3D image file format, but uses an interactive set of compressed still images that have been stitched together by special authoring software to give the illusion of a 3D interaction. In essence it offers sophisticated interactive animation. There are two types of QuickTime VR movies: panoramic movies, which allow the user to navigate a panoramic view of up to 360° from a single viewpoint; and object movies, which allow the user to view a fixed object from multiple viewpoints. In addition to this radial navigation, QuickTime VR movies can support zooming in and out, and embedded hyperlinks that can take the user either to a new QuickTime VR movie, or to a completely new web page.

⁴⁴ The latest SMIL specification and other information can be found at: <http://www.w3.org/TR/smil20>

⁴⁵ see also: <http://www.web3d.org/technicalinfo/specifications/specifications.htm>

⁴⁶ for more information see: <http://www.web3d.org/x3d.html>

Although QuickTime VR is not a genuine 3D format,⁴⁷ the intuitive navigation of photo realistic scenes and objects, coupled with the ubiquity of the QuickTime browser plug-in and the relatively low cost of authoring QTVR movies, make this an effective and appealing approach for presenting 3D cultural information on the web.⁴⁸

Other current formats for virtual reality include: MHEG-5 – an ISO standard initiative (ISO/IEC 13522: “Information technology – Coding of Multimedia and Hypermedia Information”); Java 3D – a platform-independent 3D graphics format for Java-based applications and applets;⁴⁹ MPEG-4 standard that defines how digital audio and video media can be encoded, compressed, streamed for efficient delivery over networks, and reassembled correctly for presentation.⁵⁰

4.9.3 The geophysics data is being collected only by the ADS who currently accept three file formats and have data only in two of them. There is no agreed *de facto* standard employed by various proprietary software (e.g., Geoplot, Surfer, Contours, etc.). However, there is an initiative at the University of Bradford to develop an ASCII text-based format for geophysics data. The authors of the new format are collaborating with the software developers and there is promise for the new file format to be made compatible by the relevant software packages through an export function into this format. Until the appearance of the standard format and the exporting support from software packages, the ADS is recommending (and following this advice itself) to archive the data and the control info of geophysics files separately as ASCII texts. This may mean loss in functionality of original data resource and there have been problems with separating the header part of a file.

4.9.4 The archival preservation of web resources is an increasingly burning issue for many archives around the world. The problem with web archiving is perceived not so much as the question of file formats but rather lying in capture, description and later rendering of web pages that include hyperlinks to other pages, make extensive use of various plug-ins and include formatting that is hard- or software specific. With the development of technology and new file formats, the web is becoming an increasingly popular medium for distributing text, image, sound and video resources, and the AHDS will definitely start receiving web-based data resources as deposits in the nearest future. The AHDS Service Providers did not list any explicitly web-based resources in their collections, although they preserve over a gigabyte of HTML format data.

There is no file format, as such, for preserving web resources, although XML comes close to becoming a *de facto* standard that permits to describe all the various data types that may make up a web page. The prevailing preservation strategy for web pages is to break them into their constituent elements by data type and apply preservation processing to each according to their needs. The integrity of a web page would be preserved through extensive metadata that should also describe any functionality that is at risk due to obsolescence of plug-in software or is dependent on a specific hardware device (e.g., sound card).

The longest experience with archiving and preserving web resources can be found in the library community where several national-level programmes have been set up to archive whole internet domains. There is some literature available from these projects regarding their methods of preserving web documents (see Appendix II for further details).

⁴⁷ see: <http://www.apple.com/quicktime/qtvr>

⁴⁸ Tony Gill, “3D Culture on the Web”, 2001

⁴⁹ <http://java.sun.com/products/java-media/3D/>

⁵⁰ <http://mpeg.telecomitalia.com/standards/mpeg-4/mpeg-4.htm>

4.9.5 The preservation of executable code, programs and whole packages is considered to be part of the technology preservation strategy that cannot, as a rule, rely on standard or open formats. The instructions for the computer are stored in a program in a binary code form. The executable program needs to be adequately described for its hard- and software requirements (e.g., the operating systems, compiler, etc.) and preserved together with the hard- and software required for its use and/or rendering.

The AHDS currently has only one executable data resource – a multimedia CAL program held in VADS collections. The program requires, as a minimum, MS Windows 3.1 operating system, which should be archived at VADS alongside the CAL resource itself.

4.10 CONCLUSIONS

In order to maximise access to archived data resources, open standard formats should be used when both creating and preserving digital resources. The use of open file formats helps with software interoperability, ensuring that data resources remain reusable and can be processed by a variety of applications. In some cases there may be no relevant open standards or the formats may be too new for conformant software tools to be widely available. The use of proprietary formats may be acceptable in these cases, but the formats must be supported by more than one software package (the so called *de facto* standards). A migration strategy should be explored for these formats that will enable a conversion to open standards in the future.

The AHDS Service Providers are aware of the preservation issues related to proprietary and non-standard file formats and have chosen open or industry standard formats for the preservation of most data types in their collections. Preservation formats for some data types (e.g., CAL, GIS, video) remain problematic as the software marked for these data types is still developing or dominated by a few developers. The AHDS Service Providers are monitoring the developments and (thanks to active user demand) are keeping their preservation strategies up to date with the technology development.

The difficulties related to preservation file formats that the AHDS Service Providers have identified are related more to the storage space required by some file formats and documentation of preservation processes, rather than the complexity of dealing with a specific file format. The expected new file formats do not pose any significant difficulties for the Service Providers.

The choice of preservation file formats by the AHDS Service Providers cannot be considered completely risk-free, but it could be assessed as being low-risk. The choice of preservation formats should, however, be complemented with a detailed documentation of preservation processing, where the AHDS as a whole is currently short of an adequate strategy.

5. Assessment of preservation management practices

This chapter will follow the data resources through their life-cycle within the AHDS collections and identify key stages in this cycle that may or do affect the preservation of the data resources. It will also identify file format related factors that may affect the overall cost of the preservation process — both in terms of labour (time) as well as funds.

The basic “rules of thumb” for assessing the complexity and, hence, the cost of digital preservation have been summarised as follows:⁵¹

	Simpler Lower cost archive	➔	More complex Higher cost archive
Data types & formats	Limited number		Large number
Rights	Ownership		Non-ownership
Control	High degree of control		Low degree of control

Figure 6. Assessment of cost of digital preservation processes.

The AHDS has a high degree of control over the archived data after their acquisition: the Service Providers are permitted to convert practically all of their collections into formats that they have deemed suitable for long-term preservation. The level of control over the archived data is often closely linked to data ownership, which the AHDS does not have over the data resources that it collects, archives and disseminates. However, the remit for processing of these resources that AHDS has negotiated with its depositors leaves it with the very high degree of control over the data once it has been accessioned. However, the level of control that the AHDS exercises over the file formats that the depositors create and deposit their data in, is low. It is the processing (conversion and migration) and documenting of data resources that can become costly in a low-control situation over the input and output formats from a digital archive.

Preservation service as the link between the ingest and dissemination must be managed, documented and tailored to smooth the flow of data ‘through’ the archive safely and accurately (i.e., without any significant losses). It is with the potential ‘bottlenecks’ in this flow that this chapter is concerned with.

⁵¹ from S. Granger, K. Russell, E. Weinberger, “Cost Elements of Digital Preservation”, 2000, p. 5

5.1 ACQUISITION FILE FORMATS

The nature of the AHDS's relationship with its (potential) depositors is mostly guiding, educating and recommending (with a few exceptions). This determines an open acquisition policy rather than a strict, prescriptive control, and the AHDS Service Providers have to accept data in a large number of different file formats. The list of deposit formats is essentially a compromise between file formats that the data depositors are likely to be wanting to deposit their data in and file formats that the AHDS can realistically (i.e., within feasible cost limits) preserve and make available for the future. And even these lists of accepted file formats are not exclusive and the Service Providers are prepared to consider other formats for deposit.

Such openness may potentially have its price and result in a situation where the AHDS needs to commit significant resources to conversion of legacy file formats, or very new file formats that have not become stabilised yet. Nevertheless, the openness of the list of deposit file formats should be interpreted towards accepting newer formats rather than older, legacy formats, because the software industry is nowadays driven to standardisation and compatibility much more than it has been in the past. This makes the new and emerging file formats generally more open, transparent and hence easier to archive and preserve.

The lists of deposit formats that the AHDS Service Providers are accepting should be updated regularly to reflect new, appearing formats (preferably open and standard formats, e.g., X3D). The list should also be co-ordinated among Service Providers to overlap for common data types.

5.2 PRESERVATION FILE FORMATS

This report has provided an in-depth analysis of preservation file formats that the AHDS Service Providers are using and has concluded that they are predominantly standardised formats or, in the absence of open formats, the *de facto* industry standards are used which do not pose high risks for the long-term archival preservation.

Standard formats often offer safety of future usability at a cost of loss in functionality (e.g., simple ASCII text). In this respect the AHDS preservation practices have favoured continuing access to functionality of archived digital resources.

The number of preservation formats has been kept reasonably small, i.e., it is not too restrictive and offers alternatives where several 'safe' standards or formats exist. Yet with some data types that the AHDS has thus far had little experience (e.g., digital video and audio), the list of preservation formats could be more restrictive and more standards-driven.

The file formats survey concluded that the formats chosen for preservation have so far not created any significant problems for the AHDS Service Providers and the 'track record' of preservation is good — with no loss of data or access.

5.3 CREATION OF THE PRESERVATION VERSION

The AHDS practices conversion of the ingested data resources into file formats suitable for long-term preservation and creation of a preservation version of the data resource that is retained for long term. Despite the successful preservation record, the processing of deposited data into a preservation version can not be said to be sufficiently defined nor controlled.

The file format survey and subsequent interviews with the AHDS Service Providers revealed that no formal procedures or transparent guidelines for creation of the preservation version have been defined at any of the Service Providers, and that the conversion processing is insufficiently documented. In some cases the conversion has been determined more by the cost issues than the needs of the archived data resources. The main (in some cases the only) method

of ensuring the quality of the whole conversion process is the content validation of the result — the preservation version (see next subsection).

The conversion of files from their original format into the format chosen for preservation should be as simple and straightforward as technically feasible and if possible avoid multiple conversion stages or intermediate, ‘transition’ formats. One Service Provider has been using a service provider for file conversion in which case the required quality of the conversion outcome has been fixed in the service contract, but the documentation of the conversion processes carried out is lacking.

The long-term preservation of many of the data resources in the AHDS collections will require their migration to new formats at some point. This point in time may arrive when the deposited original version of the data resource has become obsolete and unusable, making it vital that the migration decisions and processing can rely on the knowledge of previous conversions of the data resource. To facilitate this, the current AHDS practice of preservation version creation is lacking in particular in these areas:

- Formalised procedures for choosing the preservation file format for a given deposit file format.
- Description of the conversion process (e.g., conversion tools used, date and time of processing, etc.)
- Description of methods of quality assurance of the conversion process.
- Documenting the validation or quality control of the conversion outcome.
- Documenting the loss in functionality and/or significant properties of the original data resource resulting from the conversion.

The AHDS has had little experience with migrating its collections and the need for such documentation has not been clearly apparent. The proposed AHDS Preservation metadata framework is addressing these issues and offers a solution to better transparency of the process of creating the preservation version (see also Documentation below in section 5.11).

5.4 VALIDATION OF THE PRESERVATION VERSION

The validation of results when converting the original deposited data file into a preservation file format serves the purpose of quality assurance as in the future the dissemination version of a data resource may have to be created from the preservation version and any damage or data loss in the preservation version will be impossible to rectify at that point.

The file format survey asked a question regarding the validation of preservation versions and the answers demonstrate a relatively low level of quality control. The answers included:

- no validation;
- visual checks;
- sample data checking;
- automatic routines;
- checked by software prior to dissemination.

Depending on the data type and file format, the data content validation should be as rigorous and thorough as possible, and must be well documented (cf. previous subsection). Formalisation of procedures and description of processing in preservation metadata are solutions for improving the validation of preservation version creation.

5.5 COMPRESSION TOOLS

Compression algorithms can reduce file sizes considerably but in general compression adds additional complexity to the preservation process and is normally not recommended for the storage of archived digital resources. With the reducing cost of storage and increasing capacity of storage media, compression for storing is becoming less necessary. However, it can have a role in data transmission – both for acquisition and dissemination.

The file format survey identified that all AHDS Service Providers are accepting deposits in compressed formats and that some are also archived in compressed format. The list of compression tools in use is the following:

Compression software	Version
WinZip	
GNUzip	.gz
	gzip 1.2.4 Solaris
Pkzip	
Tar	tar
	Solaris 7 tar
UNIX Compression	.Z
Stuffit	
.zip	

Table 13. Compression tools used by Service Providers.

While for deposits and dissemination the file compression is acceptable practice, it should be avoided for preservation versions. If, nevertheless, the file compression has been used for preserved copies, the tools used should be documented in the preservation metadata.

5.6 STORAGE MANAGEMENT

The efficient management of archival storage is the key to low-risk and low-cost preservation service. The storage strategy must ensure the:

- creation of a sufficient number of copies of preserved data resources;
- sufficient distribution of copies (i.e., using both on-site and off-site storage);
- required level of security and access control to the archived data resources;
- fast and convenient access for saving or retrieval for authorised users of the archive.

All AHDS Service Providers are reliant to some degree on the technical infrastructure of their hosting institutions. This has both good as well as negative sides. On the positive side are:

- established back-up routines and systems;
- controlled access to servers;
- technical support;
- fast network access.

In the cases of some Service Providers also:

- extra copies of all preserved data resources;
- off-site storage of copies;
- sharing licences for software with the hosting institution;
- negotiable extensions to storage space;
- extra levels of data security (e.g., firewall).

On the downside, the Service Providers have reported:

- difficulty in gaining physical access to servers;
- limits to storage space;
- reduced control over access to storage;
- over-reliance on technical services provided by the hosting institution;
- cost.

On the whole, there has been only a few times when the listed difficulties have constituted a risky situation for the preservation or other services that the AHDS Service Providers are providing. However, the current practice with creation of safety copies, off-site storage of copies and regular checking of storage media for readability can only be considered adequate in the case of the HDS that is relying on the UK Data Archive for its preservation management. The ADS has recently started using off-site storage from a digital storage service provider for a copy of their collections. The other AHDS Service Providers are relying on either regular back-up copies or do not make extra copies of their collections for preservation. No formal policy on periodic checking of storage media for readability has been put in place which, in particular for VADS, should be essential practice.

With the exception of HDS and PADS, none of other the AHDS Service Providers has established a preservation strategy or a formalised guidance on the management of preservation copies and storage. However, there is an expectation that a centralised AHDS preservation facility or service will help to fill this gap in the future.

5.7 AUTHENTICATION METHODS

Ensuring the continuing authenticity and reliability of the archived data resources is the basis for a quality data service. It is also a sign of quality and security of the preservation service within a digital archive. Using audit trails, digital signatures and calculated checksums is the most common way of keeping a log of changes made to the data resources in archival storage. Some of the AHDS Service Providers are using MD5 checksums for authentication of their collections. These checksums can be compared at regular intervals to detect any files that have been changed since the last check. It is general good practice in digital archives and other AHDS Service Providers, too, should implement authentication methods. (See also Fixity metadata section in the AHDS Preservation Metadata Framework.)

5.8 PRESERVATION METADATA

Preservation metadata – both technical and administrative – is kept to document the preservation processes of a digital archive, make them transparent and accountable and also in preservation processing itself.

The AHDS Service Providers do not have a formal preservation metadata scheme and only some have implemented a way of recording the technical details and requirements of the archived data resources and describing preservation practices. Some rely on automated conversion scripts as documentation.

This area is in urgent need of improvement and has been covered by the AHDS Preservation Metadata Framework. See also subsections on documentation, collections management and validation of the preservation version.

5.9 CREATION OF THE DISSEMINATION VERSION

The AHDS practice is to create separate version(s) for dissemination of its collections if these are not deposited and/or preserved in file formats that are suitable for easy dissemination. The dissemination version can, in theory, be created from both – the original deposited version and the preservation version. If the original version is still usable and can successfully be converted into the required dissemination file format, it is the original deposited file that should be used. If the deposited file has become obsolete or it proves too costly to create the new version for dissemination from the original, only then should the preservation version of the data resource be used. This practice will ensure the quality of the new dissemination format (i.e., in case the preservation format included any errors or had omitted some functionality that can be preserved in direct conversion from original format to a new file format) and will safeguard the preservation files from any accidental corruption.

The AHDS Service Providers follow this best practice whenever possible, although no formal policy or guidance exists to regulate the practice. Implementation of the AHDS Preservation Metadata Framework will ensure documentation of some of the processing steps in this stage. For better storage management, it would be good practice to remove obsolete dissemination versions (that are no longer requested by users) to free storage space. However, given the relatively low cost of storage media and the small number of individual data resources in the AHDS collections, this is not perceived as a problematic issue yet. The regular monitoring of the “usefulness” of a dissemination version should, nevertheless, be built into the Technology Watch service (see next subsection).

5.10 TECHNOLOGY WATCH

The purpose of a formal process of ‘technology watch’ is to maintain a register of hardware and software capacity and the preservation metadata in the archive to prevent the risks arising from technology becoming obsolete through rapid development. An archive with little control over formats and media received and a high degree of diversity in its collections will find this function essential.⁵² Failure to implement an effective technology watch will risk potential loss of access to archived resources and higher recovery costs.

All AHDS Service Providers have established a technology watch in some form, but none of them has formalised it or set its principles out in a policy. The AHDS would benefit from an co-ordinated Technology Watch service or facility that would consider the needs of all Service Providers which are to a large degree overlapping. The AHDS could also consider establishing a Technology Watch service that would notify its users and depositors of new developments in technology and archiving practices that have a bearing on how (or whether) their digital resources can be preserved and made available for further use.

⁵² M. Jones, N. Beagrie, “Preservation Management of Digital Materials: A Handbook”, 2001

5.11 CONCLUSIONS

The preservation practices and processes implemented by the AHDS Service Providers are to a large extent following current best practice. There are, however, serious shortcomings in making the preservation processes formally regulated, transparent and well documented. The main three areas that currently are inadequately implemented within the AHDS are:

5.11.1 Collections management, which only two Service Providers have made an attempt to implement through a collections management database. One of these has not gone further than designing the database, but has not started using it. Only the ADS has designed and is using a database to record the technical details about its accessioned collections, file formats, conversion processes, file locations and checksums. Some of the other Service Providers collect and keep the same data, but have not organised it as a single management database, metadata schema or formal set of descriptions. Collections management could be used as a collective term for all ten stages of a digital resource's life-cycle that have been listed and discussed in this chapter. Establishing proper collections management would improve all of the listed ten aspects of Service Provider's work.

5.11.2 Documentation and metadata, that would describe and document the decisions made during the preservation, processes carried out and results that were obtained. Documenting the preservation decisions and processes falls largely under the administrative metadata that is included in the proposed AHDS Preservation Metadata Framework. The implementation of a consistent metadata schema achieves several goals at the same time: it helps to formalise the processes, it ensures consistency of practice, it identifies responsibility for carrying out the processes, and it makes the decision-making transparent for users both inside as well as outside the archive.

5.11.3 Preservation policy, that would provide a comprehensive statement about the preservation strategy of the AHDS collections, deal with all aspects of preservation and apply to all materials held by the AHDS Service Providers. The main aim of the AHDS preservation strategy should be to make the collected data resources accessible, while ensuring their survival and usability in perpetuity. The AHDS preservation policy should outline how the preservation service fulfils the objectives set by the strategy, how the preservation is organised and promoted as an integral part of the AHDS collections management, and how it ensures rational use of resources. At present, only PADS has made an attempt to issue a separate statement⁵³ as to the strategy of fulfilling the main mission of a Data Service: to preserve the data so it can be meaningfully used. The HDS has included a section on preservation policy in its Collections Manual.

There is considerable overlap between the AHDS Service Providers in terms of data types that they hold and file formats that they use for preservation. Yet there is little co-ordination of the efforts that they all make separately to preserve the archived data resources and significant differences in preservation practices. While there is little immediate risk to the survival of the data resources that have been deposited with the AHDS, the growing collection size will soon necessitate the homogenisation of preservation practices across all Service Providers. Taking a more collaborative approach to preservation and its requirements would be the first step towards common and adequate preservation practices.

⁵³ "PADS Digital Preservation Strategy" (<http://www.pads.ahds.ac.uk/preservation>)

6. Recommendations

Recommendations made in this chapter should be read in conjunction with the recommendations in the consultancy report “Preservation in the AHDS: A critical review” (April 2002).

GENERAL RECOMMENDATIONS

1. The AHDS should establish an over-arching preservation policy that would state the AHDS digital preservation mission, strategy and principles.
2. A preservation service or facility should be set up to hold preservation versions of all AHDS collections. As a minimum, the preservation centre must:
 - 2.1 Implement and maintain a preservation metadata schema.
 - 2.2 Establish a collection management system and maintain it.
 - 2.3 Offer sufficient storage for all AHDS Service Providers and ensure standard archival storage management principles.
 - 2.4 Offer sufficient security of archived data while providing AHDS Service Providers with flexible and easy access to their data resources.
 - 2.5 Implement a process of technology watch.
3. The AHDS should implement a preservation metadata scheme, in particular for documenting the administration of preservation management and preservation processing.
4. The AHDS should issue detailed guidance to its Service Providers on digital preservation management and practices (cf. the forthcoming “The AHDS Preservation Procedures Manual”).
5. The AHDS should co-ordinate regular updating of file format lists that the Service Providers are publishing so as to include the development of new standards and formats. The lists of acceptable file formats should be harmonised among the Service Providers for overlapping data types.

DETAILED RECOMMENDATIONS

6. The AHDS should agree on at least one mark-up language (e.g., XML, SGML) that all Service Providers could list as a preferred preservation format for textual data.
7. Migration strategies for converting RTF and PDF files into XML should be explored and evaluated.
8. The AHDS Service Providers should adhere to the good practice of recording the text encoding scheme in the metadata of the textual resource they archive. This is particularly valid for multi-lingual and historical data resources.

9. The AHDS should agree on a unified or standard method of formal description for the structure of relational databases. This description should be stored and kept as part of the structural metadata for the use of archived resources.
10. XML should be explored as a standardised format for preserving both data and descriptive elements of spreadsheet data.
11. The AHDS Service Providers should explore migration paths for their digital video and audio data resources to make better use of standard file formats, e.g., the MPEG family. Where this appears not feasible, the Service Providers should archive the originating software application alongside the data resource.
12. The AHDS should co-ordinate monitoring of developments in the digital video and audio file formats in order to amend the metadata requirements for deposited video and audio data.
13. The AHDS Service Providers should continue monitoring the developments in the CAD, GIS, Geophysics and Virtual Reality file formats and establish safe migration strategies for conversion of their collections into standard formats whenever possible.
14. The AHDS Service Providers should monitor the developments in web document archiving and make preparations for accessioning web resources.
15. The AHDS Service Providers should refrain from using file compression whenever possible for preservation versions of their collections.
16. The AHDS should co-ordinate a review of the installation sets of application software and operating systems that the Service Providers currently hold and establish a 'technology preservation' strategy to ensure that at least one copy of installation sets is retained within the AHDS network for legacy software.

RECOMMENDATIONS FOR EXISTING POLICY TEXTS

17. The AHDS Service Providers should review their use of the term 'data type' in their Collections Policies and Guides for Depositors to make it refer to the same concept.
18. The policy documents should be kept up-to-date and reflect the file formats that are currently in use by data creators and depositors. It should be avoided that depositors must convert their data resources into an older version of a format for deposit with the AHDS.
19. Depositors should be offered more explicit or detailed guidance on what data and description they need to deposit when they are only depositing a brokered or linked resource. The current Guides for Depositors do not make sufficient difference between the five levels of collection that the AHDS has identified.
20. The recommended preservation practices in the individual volumes of the Guides to Best Practice series should not differ more than the requirements of a particular data type prescribe.

21. The Guides to Best Practice and policies should be updated to reflect the appearance of new texts and provide direct links to these. The policies and Guides should be spell-checked for typing errors and should be equipped with dates of latest revision – date and version number should preferably appear on both the on-line and printable versions of texts.
22. The Guides to Best Practice and policies should use full names and titles of institutions, and projects, rather than acronyms; if possible, provide links to these as a separate appendix to the text.
23. The AHDS should consider offering its users and depositors guidance for good preservation practices for the period prior to depositing their data with the AHDS Service Provider. Educating the depositor not only in good description practices but also in proper storage, back-up, multiple copies of files, version control and conversion practices, although basic knowledge, can save both time and resources for the AHDS. The ADS Guide to Good Practice for Geophysics Data includes a section (4.3) that can serve as a good starting point for such guidance.

Appendix I File format survey letter and questionnaire form

20/07/2002

Dear colleague,

One of the deliverables that forms a part of the AHDS preservation consultancy is an overview of data types and file formats currently used by AHDS service providers. Although the previous consultants have already enquired during their interviews about the file formats in use, a more detailed account of file format handling practices will be required for completing the preservation consultancy deliverables 2 and 7. I therefore kindly ask you to take the time for filling in the questionnaire below and return it to Hamish James (hamish@essex.ac.uk) by July 29th.

The questionnaire is divided into two parts: first, a preliminary list of file formats that are used at the Service Provider, that I have compiled based on the publicly available texts and which I ask you to correct and amend; and, second a few questions detailing the file format handling and conversion practices as well as legal/copyright requirements attached to deposits.

If you have any queries or comments, please do not hesitate to contact me.

Thank you for your help,

Raivo Ruusalepp
Consultant
Estonian Business Archives
raivo@eba.ee

Data Type	Ingest/Deposit File Formats	Volume Mb/Gb	Preservation File Formats	Volume Mb/Gb	Dissemination File Formats	Volume Mb/Gb	Comments
Audio	Preferred formats:		Preferred formats:		Preferred formats:		
	Acceptable formats:		Acceptable formats:		Acceptable formats:		
Multimedia	Preferred formats:		Preferred formats:		Preferred formats:		
	Acceptable formats:		Acceptable formats:		Acceptable formats:		
Spreadsheet	Preferred formats:		Preferred formats:		Preferred formats:		
Scanned paper documentation			Preferred formats:		Preferred formats:		
Other							
Virtual Reality							
CAD							
GIS							
Compression	Tools used:		Tools used:		Tools used:		

Sources: texts where the file formats were taken from into this table for each SP.

2. If known, please specify the character encoding used for the text files.

(i.e., code pages, ISO standards, UTF-8, UCS-4, etc.)

Is any provision made for storing the information on database structures separately from the actual content (tables) of databases?

(e.g., when storing a database as a delimited text file, etc.)

3. Any other comments?

Section II

The following questions aim to clarify the practices of creating and managing file formats in the work of your data service. Please adjust the size of text boxes below to use as much additional space as you require to answer the questions.

4. What is the choice of file formats for preservation based on?

(e.g., a fixed procedures manual, technology watch, risk analysis, committee decision, cost analysis, long-term preservation requirements, etc. – please expand and attach copies of any relevant texts.)

5. Does the Service Provider have any detailed guidelines in place for converting one file format into another?

(e.g., when creating a preservation format from a deposit file format; during migration; etc. – please attach copies of or references to such texts.)

6. Who is responsible for initiating (requesting) the creation and the actual creation of the preservation and dissemination file formats at the Service Provider?

(e.g., deposit officer, preservation officer/service, user services, etc.)

7. Are the newly created file formats checked for completeness and consistency, or validated in some other way against the original files?

8. Are there any file formats among your collections that are specifically required to be kept in their original format by the depositor?

(e.g., are there any files that have specific authenticity requirements or that need to maintain certain functionalities? – please list all such file formats that may potentially be problematic for long term preservation.)

9. Are there any cases where it has proved impossible to preserve the original deposited format? What was the file format?

10. Are there any file formats that were not listed in the table above, but that the Service Provider is expecting to have to deal with in the near future? Have there been any decisions made yet as for the preservation format of these, or any preparations made for accessioning these file formats?

11. Do you see any of the listed file formats as problematic for the PADS from the preservation point of view? Are there any file formats you would rather not have to deal with? If you could change something about the preservation file formats, their creation or management, what would it be?

Thank you for your time!

Please return the questionnaire to Hamish James at hamish@essex.ac.uk.

Appendix II

Bibliography and useful sources of information

Samuel Brylawski, "Preservation of Digitally Recorded Sound", in: "Building a National Strategy for Preservation: Issues in Digital Media Archiving", Council on Library and Information Resources / Library of Congress (April 2002)
<http://www.clir.org/pubs/reports/pub106/sound.html>

"Cedars Guide to Digital Preservation Strategies", CEDARS Project (2002)
<http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html>

James Coleman, Don Willis, "SGML as a Framework for Digital Preservation and Access", CLIR (1997)

DLM-Forum, "Guidelines on best practices for using electronic information" (1997)

Charles Dollar, "Authentic Electronic Records: Strategies for Long-Term Access", Cohasset Associates (1999)

Michael Ester, "Digital Image Collections: Issues and Practice" CLIR (1996)

Tony Gill, "3D Culture on the Web" // *RLG DigiNews*, vol. 5 (2001), no. 3
<http://www.rlg.org/preserv/diginews/diginews5-3.html#featured>

Stewart Granger, Kelly Russell, Ellis Weinberger, "Cost Elements of Digital Preservation", version 4.0, CEDARS Project (October 2000)
<http://www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc>

Guidelines for keeping records of web-based activity in the Commonwealth Government (March 2001) http://www.naa.gov.au/recordkeeping/er/web_records/archweb_guide.pdf

Maggie Jones, Neil Beagrie, "Preservation Management of Digital Materials: A Handbook" (2001)

A. Kenney, N. McGovern, P. Botticelli, R. Entlich, C. Lagoze, S. Payette, "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism" // *D-Lib Magazine*, vol. 8 (January 2002), no. 1 <http://www.dlib.org/dlib/january02/kenney/01kenney.html>

G. Lawrence, W. Kehoe, O. Rieger, W. Walters, A. Kenney, "Risk Management of Digital Information: A File Format Investigation", CLIR (June 2000)
<http://www.clir.org/pubs/reports/pub93/pub93.pdf>

Kevin McDowell, "Export a Word Document to XML" (May 2001)
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnword2k/html/odc_expwordtoxml.asp

Nancy McGovern, "Cornell University Electronic Student Records Systems Project Report" (2000) <http://rmc.library.cornell.edu/online/studentRecords/default.htm>

John Ockerbloom, "Archiving and Preserving PDF Files" // *RLG DigiNews*, vol. 5 (2001), no. 1 <http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2>

Steve Thomas, "File Formats for Electronic Text" (2002)
<http://www.library.adelaide.edu.au/~stthomas/papers/etext-formats.html>

"To Preserve and Provide Access to Electronic Records", TemaNord 1996:549, Nordic Council of Ministers (1996)

UNESCO Memory of the World Programme, "Recommendations of the Committee on Technology for Consideration by the International Advisory Committee" (1995)

Howard Wactlar, Michael Christel, "Digital Video Archives: Managing Through Metadata", in: "Building a National Strategy for Preservation: Issues in Digital Media Archiving", Council on Library and Information Resources / Library of Congress (April 2002)
<http://www.clir.org/pubs/reports/pub106/video.html>

File Format Extensions Encyclopaedia
http://fileformat.virtualave.net/ext/ext_a.htm

The Unofficial TIFF Home Page
<http://home.earthlink.net/~ritter/tiff/>

A selection of (unofficial) file format definitions
<http://myfileformats.com/index.php>

PADI: Preservation of networked digital material
<http://www.nla.gov.au/padi/topics/45.html>

Kulturarw3: Long time preservation of electronic documents
<http://www.kb.se/kw3/ENG/Default.htm>

PANDORA - Preserving and Accessing Networked Documentary Resources in Australia
<http://pandora.nla.gov.au>

Appendix III

A selection of standards and definitions of file formats

Textual data

ISO 646:1991 ISO 7-bit coded character set for information interchange
ISO 8859 Information technology – 8-bit single-byte coded graphic character sets (Parts 1-16, 1998-2001)
ISO 10646:2000 Universal Multiple-Octet Coded Character Set (UCS)

ISO 8879:1986 Standard Generalized Mark-up Language (SGML)
ISO 15445:2000 HyperText Mark-up Language (HTML)
ISO TR 22250-1:2002 Regular Language Description for XML (RELAX) – Part 1: RELAX Core

PDF File format specification, ver. 1.4
<http://partners.adobe.com/asn/developer/acrosdk/docs/filefmtspecs/PDFReference.pdf>

Database data

ISO 9075 Information technology – Database languages – SQL (Parts 1-13, 1999-2002)
ISO 11179 Information technology – Specification and standardization of data elements (Parts 1-6, 1994-2000)

Image data

ISO 15444 Information technology – JPEG 2000 image coding system (Parts 1-4, 2000-2002)

ISO/TS 12029:2002 Forms design optimization for electronic image management
ISO/TS 12033:2001 Guidance for selection of document image compression methods
ISO/TR 14105:2001 Human and organizational issues for successful Electronic Image Management (EIM) implementation

TIFF 6.0 File format specification
<http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf>

PNG File format specification
<http://www.w3.org/TR/REC-png.pdf>

Digital video data

ISO 11172:1993 Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s (MPEG-1) (Parts 1-5, 1993-1998)
ISO 13818-2:2000 Information technology – Generic coding of moving pictures and associated audio information: Video (MPEG-2)
ISO 14496-2:2001 Information technology – Coding of audio-visual objects – Part 2: Visual (MPEG-4)

MPEG-7 Overview

<http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>

QuickTime File format specification

<http://developer.apple.com/techpubs/quicktime/qtdevdocs/QTFF/qtff.html>

Digital audio data

ISO 11172-3:1993 Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio (MPEG-1)

ISO 13818-3:1998 Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio (MPEG-2, Layer 3; MP3)

ISO 14496-3:2001 Information technology – Coding of audio-visual objects – Part 3: Audio (MPEG-4)

RIFF WAVE File format specification

<http://www.techweb.rfa.org/docs/RIFFWAVE.htm>

Ogg Vorbis I File format specification

<http://www.xiph.org/ogg/vorbis/docs.html>

CAD data

AutoCAD R13/R14/R2000 DWG File format specification (version 2.0)

<http://64.85.21.186:90/downloads/spec/formatSpec13-15.rtf>

Standard for long term CAD data archiving

<http://www.mosla.org/download/main.html>

Vector Product Format specification

<http://164.214.2.59/publications/specs/printed/VPF/vpf.html>

SMIL File format specification

<http://www.w3.org/TR/smil20>

GIS data

ANSI INCITS 320:1998 Information Technology - Spatial Data Transfer Standard (SDTS)

BS 7567:1992 Electronic transfer of geographic information (NTF) (Parts 1-3)

ESRI Shapefile Technical Description

<http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>

MapInfo MIF/MID File format specification

http://www.mapinfo.com/free/docs/mipro/mipro_70_users.pdf (see Appendix J)

Other

ISO 14772-1:1998 Information technology – Computer graphics and image processing – The Virtual Reality Modelling Language – Part 1: Functional specification and UTF-8 encoding

ISO 13522-1:1997 Information technology – Coding of multimedia and hypermedia information (Parts 1-8, 1997-2001) (MHEG)

Extensible 3D (X3D) File format specification

<http://www.web3d.org/technicalinfo/specifications/specifications.htm>

Storage media

ISO 9660:1988 Information processing – Volume and file structure of CD-ROM for information interchange

ISO 10149:1995 Information technology – Data interchange on read-only 120 mm optical data disks (CD-ROM)

ISO 12142:2001 Media error monitoring and reporting techniques for verification of stored data on optical digital data disks

ISO 10922:2000 Information on Optical Disk Cartridges (ODC) shipping packages and ODC labels

ISO 16448:2002 Information technology – 120 mm DVD – Read-only disk

ISO/TR 12037:1998 Recommendations for the expungement of information recorded on write-once optical media

ISO/TR 12654:1997 Recommendations for the management of electronic recording systems for the recording of documents that may be required as evidence, on WORM optical disk

ISO/TR 10623:1991 Requirements for computer-aided design and draughting – Vocabulary

BS 4783 Storage, transportation and maintenance of media for use in data processing and information storage (Parts 1-8, 1988-1994)