*erpa*seminar

File Formats for Preservation
Wien, 10-11 May 2004

FINAL REPORT

File Formats for Preservation

Austrian National Library, Wien

10-11 May 2004

Österreichische Nationalbibliothek

DCC

Table of Contents

Introduction

File formats are a crucial layer, indeed a hinge between the bits in storage and their meaningful interpretation. The proper access to and display of content depends entirely on the ability to decipher the respective bitstream, and therefore on precise knowledge about how the information contained within is represented. Consequently, file formats are one of the core issues of any digital preservation approach, and file format obsolescence is a major challenge for anybody wanting to preserve digital files.

Participants from nineteen countries, archivists, librarian, computer scientists, and others, gathered in Vienna on 10/11 May to discuss file format issues concerning digital preservation. Eleven presentations offered background concepts and examples to the audience, while discussions and breakout groups served to question the concepts presented. During breaks, lunch, and the seminar dinner there were further possibilities for networking and information exchange.

Venue and Organisation

The seminar was split into two major parts. On the first day, papers provided information on an introductory and general level, while on the second day file format questions were treated according to particular areas, namely text, image and audio-visual formats.

Programme

Monday 10th May 2004

08:40	*Registration and Coffee*	
09:40	Welcome addresses	Johanna Rachinger (Director General, Austrian National Library) Seamus Ross (Director, ERPANET)
	SESSION ONE Introduction and Generalities	
10:00	The Role of File Formats in Digital Preservation: Opportunities and Threats	Frank Möhle (Swiss Federal Archives, Switzerland)
11:00	File Format Registries and the PRONOM Service	Adrian Brown (National Archives, UK)
12:00	*Lunch Break*	
13:15	*Guided tour through the Hall of State*	
14:00	The Typed Object Model: Support for Diverse Formats	John Mark Ockerbloom (University of Pennsylvania, USA)
14:40	ISO Standard Development – Overview of the Draft PDF/A Standard	Susan J. Sullivan (National Archives and Records Administration, USA)
15:20	*Break*	
15:40	Evaluating and Comparing Preservation Strategies	Andreas Rauber, Carl Rauch (Vienna University of Technology, Austria)
16:20	General Discussion	
17:00	*Adjourn*	
19:30	*Seminar Dinner, hosted by ERPANET*	

Tuesday 11th May 2004

	SESSION TWO Practical Experiences with Office, Image, Audio, and Video Formats	
09:00	Practical Experiences of the Digital Preservation Testbed: Office Formats	Jacqueline Slats (Nationaal Archief, The Netherlands)
09:40	XML as a Preservation Strategy: Experiences with the DiVA document format	Eva Müller (Uppsala University Library, Sweden)
10:20	Demonstration of the DiVA Project	Uwe Klosa (Uppsala University Library, Sweden)
10:40	*Break*	
11.00	An Overview on Image Formats	René Van Horik (NIWI-KNAW, The Netherlands)
11:40	The Flexible Image Transport System (FITS) in Astronomy	Preben Grosbøl (European Southern Observatory, Germany)
12:20	*Lunch break*	
13:20	Digitisation – The Only Viable Way to Preserve Audio Recordings in the Long Term	Dietrich Schüller (Austrian Academy of Science, Austria)
14:20	Format and File Issues for Video Archiving	Franz Pavuza (Phonogrammarchiv, Austria)
15:00	*Break*	
15:20	Breakout discussions	
16:30	Reporting and wrap up	
17:00	Closing remarks	
17:15	*End of seminar*	

SESSION ONE

Introduction and Generalities

Papers presented during session one discussed questions on a general level. Together, they attempted to give participants an introduction into the issues associated with file formats, and they presented methods, approaches, and resources that offer a high-level, but practical access to preservation file format questions.

The Role of File Formats in Digital Preservation: Opportunities and Threats

Frank Möhle, of the Swiss Federal Archives (SFA)[1], gave an introduction into file formats. While a file is nothing more than a sequence of bits, a file format is a definition of how to interpret this bit stream. Typically, file formats comprise a header, where metadata are stored, and the actual primary data. It should be noted that a file format can store several data formats. For instance, the data formats HTML, XML, SQL, and others are all stored in a file of text format. Möhle also mentioned two fundamental classes of file formats, namely text files, that can be viewed and edited using a text editor, and binary files, that usually request appropriate and specific software to be accessed in a useable and meaningful way.

Digital preservation has to guarantee the integrity, understandability, originality, authenticity, and accessibility of digital records and data. To enable this, preservation file formats have to fulfil a number of requirements. Their syntactical and semantical specifications must be public, they must be free of patent and license fees, and ideally they are standardised by a recognised standardisation body. Wide use and acceptance improve their long-term perspectives. Preservation formats must be free of any cryptographical and compression techniques, their specification should be self-contained, and they should be storage media-independent. It becomes clear from the above that, generally speaking, open formats are to be preferred over proprietary ones, since they allow for unlimited use without license fees or patent issues, and the fully available documentation eases their future handling.

Since even the best, open and widely accepted file formats are not immune against technological obsolescence; migration to other, newer formats will be a necessity during the preservation process. Thereby, information is transferred from one hardware and/or software configuration to another or from one generation of computer technology to a subsequent generation. Migration offers a number of opportunities. Since migrated documents are in an up-to-date file format, they are workable documents, and they can be accessed with current software tools, thereby lowering the user education need and the support costs. At the same time, the migration process exploits advances in technology. Finally, continuous migration also entails refreshing of the support media.

However, there are some threats involved as well. Migration is not an established, uniform process, but always a highly specialised transformation of data, involving considerable human and financial input at an unpredictable rate. Also, every conversion carries the risk of data corruption, and subsequent migrations increase this risk. This holds especially true for cryptographic techniques such as digital signatures. Finally, the ever growing number of file format becomes all the more difficult to track. Despite of all these threats, Möhle concluded that migration is at present the only workable solution to preserve digital files for the long term.

[1] http://www.bundesarchiv.ch/.

In the second part of his presentation Möhle sketched some leads for the discussion of several kinds of file formats, namely office documents, databases, images, and graphics. These were informed by the Swiss Federal Archives' experience with preservation formats. Most of these leads were further explored on the second day of the seminar.

Currently used office formats mostly do not meet the preservation requirements specified above. Möhle outlined different migration options, specifying that the SFA currently accept the combination of PDF and TIFF as acceptable file formats for office documents.[2]

Speaking of databases, he stated that currently no actual preservation format can be proposed. However, he presented the SIARD (Software Invariant Archiving of Relational Databases) tool developed by the SFA that allows to archive the functionality and the data of relational databases. SIARD creates three different data sets from the original database, namely an SQL DDL (Data Definition Language) file which contains the database logic and definitions for tables and views, flat files which contain the actual data sets, and an XML file for metadata. Using open standard formats and UTF-16 encoding, this preservation method allows reconstructing the database at any later point in time.[3]

Finally, Möhle introduced the two main categories of images, namely raster (or pixel) images, and vector graphics. While raster images are built up by individual picture elements, vector images are composed of geometrical figures. Therefore, they are smaller, and a number of operations can be performed on them relatively easily. For raster images, the question of compression becomes relevant. To save storage space and access time, many current raster image formats allow for compression. For preservation purposes, however, compression is not acceptable, or at the highest if it is lossless. Möhle showed an example of a lossy compression, namely a JPEG file, where the corruption of one single bit completely changed the whole picture. At the Swiss Federal Archives, only the TIFF v6 format without any compression is accepted for colour and gray-scale images, whereas for pure B/W images compression is allowed and the Fax G4 (CCIIT4) format is mandatory.

After this introductory paper a lively discussion arose. Participants with a library background challenged the focus on migration, stressing the fact that they have to preserve documents in their custody "as is", i.e. in their original format. The Universal Virtual Computer UVC has been suggested as a means to reach this. Frank conceded that the archival and library environments are different, and that, as a national archive, the SFA are in a position to at least partly influence the choice of formats they are confronted with.

In some questions and remarks the importance of guaranteeing a document's authenticity and integrity over time and the role of file formats to this task have been alluded to. While this subject kept recurring through the whole of the seminar, it was pointed out that this question surpassed the scope of the seminar, since there are a number of other factors that must contribute to authenticity.

Asked about the interval between migration cycles Frank underlined that this depends on different factors, but that at the SFA the maximum interval currently planned is ten years.

---

[2] Answering a question after his presentation, Möhle specified that PDF cannot be considered a preservation format in the strict sense of the term, and that only TIFF is accepted as an actual preservation format. He expressed his hope that PDF/A would be able to step in for current PDF soon. See on PDF/A the presentation by Susan Sullivan below.

[3] For further information about SIARD please refer to the presentation of Stephan Järmann and Stephan Heuscher at the ERPANET Workshop on "Long-term Preservation of Databases", Bern, April 2003, available from www.erpanet.org.
This short overview having met with considerable interest, Möhle kindly agreed to give a short presentation on SIARD after the end of the seminar to those wishing to get further insight.

File Format Registries and the PRONOM Service

The ERPANET Training Seminar on File Formats offered participants a close look on the promising subject of file format registries. **Adrian Brown**, of the UK National Archives, introduced this topic and illustrated it with a recent working example of a file format registry, namely the National Archives' PRONOM service.[4]

Brown defined file format registries as authoritative and publicly available sources of technical information, supporting identification, accession, preservation, and access of files. Most importantly, file format registries are expected to contribute substantially to the automation of all of the above. They are expected to be persistent, trustworthy, and publicly discoverable. It is evident that to fulfil their role registries depend on organisational permanence. This includes, but is not limited to, secure funding, available expertise, and IT infrastructure. In order to identify their content registries must use a unique identifier system. For PRONOM, the National Archives have adopted a unique identifier in e-GMS v.3 (eGovernment Metadata Standard) and are planning to implement this as a URI.[5]

When dealing with different file formats, defining the relations between them requires attention. This holds true for issues such as relationships between file formats, e.g. XML (eXtensible Markup Language) and the XML-based SVG (Scalable Vector Graphics), but also for the granularity of single formats, i.e. different versions and extensions.

Only very little information concentrated within a file format registry is in the public domain. More frequently, it is protected by copyright and either publicly available or proprietary. In the latter case, this information may be available for a fee, or can be revealed through reverse engineering as far as the license statement allows this. File format information is therefore drawn from a multitude of sources. Parts of it come from the developers, and Brown stated that some are very happy to collaborate with the registry. Other sources include other registries, digital repositories, and computer museums. Finally, an online submission form is at the disposition of those enthusiasts that wish to contribute information.

Brown then proceeded to give a quick insight into the working of the PRONOM system. It is primarily focused on software. Some key features he underlined are the provenance information that contributes to establish trust in the information, and the list of formats popular software is able to read. As a side effect, this enables using PRONOM to define migration pathways.

Finally, Brown put the National Archives' Initiative into a broader context. He mentioned the Digital Library Foundation's Global File Format Registry Project[6], which he characterised as probably the main initiative in this field, the Typed Object Model presented by John Ockerbloom in the following paper, and the Preservation Manager of the Dutch Royal Library,[7] and said that the UK Digital Curation Centre[8] was looking into the issue of file format registries as well. Brown stated that the large task ahead was probably better addressed by a network of distributed registries. Future developments are expected to include expanded coverage to other technical components, to improve standards and legislation, e.g. through escrow agreements or legal deposit, and to more automation in identification, validation, metadata extraction, technology watch, and migration pathways.

[4] See the National Archives' website at http://www.nationalarchives.gov.uk/. The website for PRONOM can be found at http://www.records.pro.gov.uk/pronom/.
[5] Uniform Resource Identifier; see http://en.wikipedia.org/wiki/Uniform_Resource_Identifier.
[6] See http://hul.harvard.edu/gdfr/. Unfortunately, no representative of the GDFR project was available to give a presentation at this seminar.
[7] See http://www.kb.nl/kb/hrd/dd/dd_onderzoek/preservation_subsystem-en.html.
[8] See http://www.dcc.ac.uk/.

File format registries are a useful instrument for preservation. But, one has to consider that at least 95% of the files to be preserved consist only of a couple of well known formats.

To identify the features a file format (e.g. a word-processing format) uses during a validation procedure would as well be of importance

The Typed Object Model: Support for Diverse Formats

Supplementing the insight into file format registries and the PRONOM system by Adrian Brown, **John Mark Ockerbloom** of the University of Pennsylvania Library[9] presented an object-oriented model to deal with file formats. The Typed Object Model (TOM)[10] focuses on functional aspects of file formats, asking what can be done with the format. To this goal, it views a format as a type combined with a sequence of encodings that represent it. A type describes the information contained within an object, namely attributes, operations, and semantics. Different encodings provide a syntax for this information. Organising these types hierarchically, with supertypes and subtypes, and providing for inheriting type information, greatly facilitates dealing with (unknown) file formats.

In practice, TOM is implemented as a distributed system of so-called “type brokers” without a central authority. These maintain and interpret the format information and can be characterised as lightweight format registries. Clients can get format information from these brokers. Moreover, they offer the plain advantages of the object-oriented system. Faced with an unknown file format, users can simply have a type broker describe it, naming attributes and operations associated with the format. More importantly, they can try to understand it better through its family tree. They may be familiar with its supertype, with one of its subtypes, or with a similar format. The type broker delivers this information, and also does format conversions, conserving particular type aspects.

The system of type brokers, organised in a peer-to-peer network, is able to originate file format definitions and to maintain them.

TOM originally was Ockerbloom’s PhD thesis in Computer Science at Carnegie Mellon University.[11] This accounts for some of the system’s limitations. Its target public clearly are computer scientists, and a background in object-oriented design is necessary to fully understand the system. Since it is heavily focused on automation, human readable documentation is not part of it. Most noteworthy, TOM originally had nothing to do with digital preservation. However, it makes important contributions to digital preservation by identifying and verifying formats, by converting formats, and by taking advantage from other information sources.

For the future, the format registry aspect of TOM is planned to be stressed. Ockerbloom also added that the current prototype of the Global Digital Format Registry[12] interoperates with TOM. This opens the field for exploiting numerous collaborative advantages.

[9] The University of Pennsylvania Library website is at http://www.library.upenn.edu/.
[10] The TOM website can be found at http://tom.library.upenn.edu/. This offers documentation on TOM, a conversion service, a type browser, a recent format registry demonstration (“Fred”), and open source software for running a TOM type broker.
[11] The Carnegie Mellon University School of Computer Science can be found at http://www.cs.cmu.edu/.
[12] See note 9 above.

One participant asked about TOM's suitability for format validation. Ockerbloom said that in principle it could be used to this goal, but that JHOVE[13] is surely more sophisticated and more suitable for this goal.

ISO Standard Development – Overview of the Draft PDF/A Standard

Susan J. Sullivan, of the US National Archives and Records Administration (NARA)[14], gave an introduction both into file format standardisation procedures and the hopes raised by PDF/A.[15]

Adobe's Portable Document Format PDF has gained wide acceptance as a de facto-standard during the last few years. In consequence, large bodies of information are maintained in PDF, and it is therefore increasingly used as a preservation format. However, PDF itself is not suitable as an archival format, since a number of its characteristics do not adjust to preservation requirements. It is owned by Adobe Inc., and, while the company has a long record of making the specification publicly available, it has no obligation to do so for future versions. PDF documents can include features that are incompatible with preservation, mainly encryption and embedded files. Also, PDF documents are not necessarily self-contained, but partly rely on system fonts and other external components. Finally, since there are multiple PDF tools on the market, there is some inconsistency with the format.

It is obvious that a stable long-term solution is needed to overcome these shortcomings if PDF is to be accepted as a suitable preservation format. In early 2002, US governmental and private institutions joined forces to initiate work towards an ISO standard that is expected to ensure preservation of PDF documents over extended periods of time, and to further ensure that PDF documents will be rendered with consistent and predictable results in the future. An international Joint Working Group was set up with support from different ISO technical committees. At the time of the seminar a second committee draft was being prepared and would be submitted shortly to National Bodies. The PDF/A roadmap provides the standardisation process to be finalised in late 2005.

Work on the PDF/A standard started from Adobe's PDF 1.4 Reference.[16] The standard will categorise the PDF components into mandatory, recommended, and prohibited, thereby defining two levels of conformance, full and minimal. It will attempt to maximise device independence, self-containment, and self-documentation.

After these introductory remarks Sullivan went on to discuss the Draft PDF/A Standard thoroughly. Some points deserve particular mention:

First of all, the scope is defined. The standard is applicable to documents containing combinations of character text, raster images, and vector graphics. After listing normative reference, terms and definitions, notation, and conformance levels, the standard expands on the technical requirements. Encryption and embedded files are prohibited. All referenced fonts must be embedded (whereby font subsets are recommended, i.e. embedding only the characters actually used), and fonts not legally embeddable are prohibited. Annotations such as sound or movie annotations, and hidden annotations, are not allowed either. There are also restrictions on actions external to the document. Metadata can be embedded using the Adobe eXtensible Metadata Platform XMP that is openly documented and available. Two informative

[13] JSTOR/Harvard Object Validation Environment. See the JHOVE homepage at http://hul.harvard.edu/jhove/jhove.html.
[14] http://www.archives.gov/.
[15] All relevant information about the PDF/A standardisation process and the PDF/A committee can be found at http://www.aiim.org/standards.asp?ID=25013.
[16] The PDF 1.4 reference is accessible from Adobe's PDF specifications page at http://partners.adobe.com/asn/tech/pdf/specifications.jsp.

Annexes will provide a summary of prohibited PDF features as well as best practices for PDF/A.

During discussion, several participants added caveats and pointed to limitations. In particular, the document focus of PDF and PDF/A was highlighted as a certain limitation. After all, there is no security that in ten or twenty years page-like documents will continue to be the prevailing format for information. Also, everything that does not fit on a page-like document will be lost in PDF/A. Finally, XML was underlined as a better alternative to PDF, though XML is still far from being appropriately implemented and used for office document formats.

Sullivan acknowledged all these points, but stated that PDF/A was mainly conceived to offer a preservation perspective to the already present PDF format. She also pointed out that Adobe insists that the PDF specification is openly documented and free, enabling everybody to write suitable software. Also, future versions of the Adobe Acrobat software and other PDF software applications might include a PDF/A mode, enabling direct generation of PDF/A documents without the detour through common PDF. Since different software vendors participate in the creation of the standard, it is realistic to expect such applications. Of course, as different participants as well as Sullivan acknowledged, the success of PDF/A will finally depend upon creators' willingness to actually use it.

Evaluating and Comparing Preservation Strategies

Andreas Rauber and Carl Rauch of the Vienna University of Technology[17] presented a decision support system for choosing the right preservation approach. Given collections with very diverse file formats and a number of preservation approaches that could possibly be used, they aim at picking the best overall approach. They employ a method called Utility Analysis and developed in the 1970s, mainly for infrastructure projects such as dams, bridges, and neighbourhoods. However, there are parallels with preservation, and the method can be adapted to fit preservation requirements.

The Utility Analysis employs an eight-step method to set objectives, evaluate alternatives, and define preferences and decide. This was presented at the example of eBook preservation. In a first step, all relevant aspects of preservation are represented in a very generic, top-down approach, resulting in a tree of objectives. For digital preservation decision support, this tree has as main branches file characteristics, process characteristics, and costs; and these are further broken down into, e.g. appearance objectives such as letter size, paragraph separation, etc. In a second step, these objectives are assigned effects, either measurable or subjective. It is possible to state non-acceptable values.

After this preparation, alternatives can be evaluated. A set of alternative methods to be analysed has to be prepared, such as migration to a newer version, migration to another, possibly open format, emulation, or no preservation effort at all. The alternatives' performance is then measured, using either original files or a testbed. Finally, the measured values are transformed to a range of numbers from one to five to make them comparable.

Defining preferences involves weighting each leaf of the tree. The sum of all leaf weights of every branch is 1. Part values per objective are calculated by multiplying leaf weights with the transformed measured values, and the sum of all part values corresponds to the total value per alternative. Once the alternatives are ranked by total value, whereby not acceptable alternatives are ranked worst, a final sensitivity analysis

[17] http://www.tuwien.ac.at/. See also http://www.ifs.tuwien.ac.at/ifs.

has to be performed, involving non-measurable influences such as specific expertise or vendor contacts.

It is evident that the construction of the tree of objectives and the weighting of objectives have the most influence on the Utility Analysis and depend strongly on the collection's requirements. Andreas and Carl envisage that a few standard trees may evolve over time for specific scenarios. One of the project's next steps will be to build and evaluate various objective trees for different preservation settings. Others will include an exhaustive listing of file format characteristics, the development of a user interface for the objective definition, and the construction of a decision support system.

After the presentation, one participant asked whether the system provided for analysing subsequent migration steps. The speakers answered in the negative; in fact, the main objective is to survive the next ten to fifteen years, then a new analysis will have to be made.

During discussion it became further clear that the system presented has a somewhat educational character in that it makes the decision process transparent, helps to identify challenges, and prioritises them. It has also been pointed out that defining characteristics and testing will be very extensive and laborious; yet this is unavoidable, for the definition of characteristics and their weighting must be conducted very carefully so as to avoid random results.

Nevertheless, the approach to use decision support systems is very promising; it opens the way to automate preservation workflows, which may be the only way to tackle the huge mass of diverse digital objects to be preserved now and in the future.

SESSION TWO

Practical Experiences with Office, Image, Audio and Video Formats

Building on the general background presented during the first day, the seminar's second part focused on specific formats, highlighting in particular experiences made within projects and research communities, and presenting best practice. The time limits prevented the discussion of a wide range of formats. Therefore, papers examined a number of widespread and common kinds of formats, namely office formats, image formats, and audiovisual formats.

Practical Experiences of the Digital Preservation Testbed: Office Formats

From 2001 to 2003 the Dutch government ran a research project to ensure the lasting accessibility and reliability of government information in the digital era, the Digital Preservation Testbed.[18] A multidisciplinary team with an archives perspective performed experiments on three different preservation strategies, namely migration, conversion to XML, and emulation (with a particular focus on the Universal Virtual Computer), and on four kinds of documents, namely text documents, spreadsheets, electronic mail, and databases.

Jacqueline Slats from the Nationaal Archief of The Netherlands, the Testbed project leader, gave an overview on practical experience of the project with office formats, mainly text documents. After an introduction to the project and to its twelve-step experiment process she presented the five basic requirements for preservation: context, content, structure, appearance, and behaviour. These were then examined with regard to text documents. It results, that the context (in particular the organisational context), all of the content, and the structure of a text document must be preserved. On the other hand, the appearance does not necessarily have to be preserved, as long as the new appearance does not alter the meaning of the original document, neither does active behaviour with the exception of the description of active links and prove of behaviour driven content.

The Digital Preservation Testbed's experiments on text document preservation include different migration approaches as well as conversion to XML.[19] Results show that migration from an older version of an application to a newer one is only suitable for the short term, while migration to the standard format PDF yields good result in representing text documents authentically, especially considering appearance. Migration of documents from one word processor to another one met authenticity requirements only after manual intervention. Finally, XML is able to represent context, content, structure, and behaviour authentically, while in order to represent appearance, an additional stylesheet is required.

Slats further expanded on both XML and PDF. As a result of the Testbed's experiments, a decision table for preservation of text documents could be presented. For documents with an implicit structure, PDF is recommended, while for documents with an explicit structure, either PDF or XML are suitable. For preservation horizons of up to ten years, a possible alternative is to rely on backwards compatibility of the file format.

The discussion following the presentation centred on different aspects of PDF and XML. It was considered unfortunate that the Dutch law explicitly mentions PDF as a

[18] The Testbed's website can be found at http://www.digitaleduurzaamheid.nl/.
[19] On the Testbed's position versus migration and XML see also their respective white papers, http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Migration.pdf and http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf.

preservation format, since the preservation perspectives of PDF are deemed insufficient. It was also cautioned against an “XML hysteria”, for there are already XML formats with non-standard compliant extensions around. This holds true both for the XML formats proposed by Microsoft and OpenOffice.org , which are not entirely W3C compliant and entail a certain dependency.

XML as a Preservation Strategy: Experiences with the DiVA document format

Started in September 2000, the DiVA (Digitale Vetenschapeliken Arkiving – Digital scientific preservation) project currently comprises ten universities in three Scandinavian countries. It aims at providing technical solutions and a functioning workflow for full text digital publishing, archiving, and dissemination of theses and other research papers. **Eva Müller**, of Uppsala University Library, Sweden, presented the DiVA project and the document format it uses.

The DiVA workflow provides producing the original document in a word processor format, following a template. The DiVA manager then converts this document into the DiVA format and stores it in a local repository. A long-term storage package is further sent into local long-term storage. The DiVA publishing system makes it possible to reuse and enhance the data directly from the source document originally created by authors, both for metadata and a digital master for electronic & printed versions; to assign a persistent identifier, store & checksum all files in a local archive; and to send a copy to the national library archives and other interested parties.

DiVA starts from the assumption that the storage format is essential. The level of enabled services depends on the granularity level of the data structure stored within the system, and the level of guarantees given for future use and understanding depends on the format used. The DiVA Document Format (DDF) is an internal format based on XML, published for, but not limited to academic publications. The project team opted for a customised format to guarantee self-description, clear structure, export support, compatibility with different metadata formats, and easy re-use of data. The choice of XML was facilitated by its open and established notation, by its support for international character sets, and by its simple and human readable text format. These characteristics all facilitate future data migration and add to the documents longevity perspectives. The DiVA project decided to use XML schema, because it provides a means for defining the structure, content, and semantics of XML documents, because it is written in XML, and because it supports data types, self-defined data types, and namespaces. Globally, a DiVA document is a metadata description of the publication, which may contain the fulltext document, and this even in a proprietary format.

Demonstration of the DiVA Project
Uwe Klosa
(Uppsala University Library, Sweden)

An Overview on Image Formats

After this first focus on text formats, **René Van Horik**, of the NIWI-KNAW, The Netherlands, introduced image formats. Quoting Murray & van Ryper, he stated that “graphics files can be considered as files that store any type of persistent graphics data (as opposed to text, spreadsheet, or numerical data, for example), and that are intended for eventual rendering and display”. The high number of different graphic file

formats is explained by the fundamentally different types that exist (raster, vector, or latent image data), by privacy and user control concerns, and by the wide range of design principles that can be followed, mainly speed and memory. Van Horik explained the frequently used raster image formats and concluded that the TIFF-format is still the most frequently used format for digital master images which addresses best preservation requirements. He examined as well three different methods to express raster images with XML: using the bit stream syntax description language (BSDL), the universal virtual computer (UVC), and the formal language for audio-visual

The Flexible Image Transport System (FITS) in Astronomy

Preben Grosbøl, of the European Southern Observatory in Garching, Germany,[20] presented an area where one single, standard, open format has been universally accepted and used for over 20 years. In astronomy, data collections have a long history. A "digital avalanche" has massively increased the amount of data to be stored and analysed. Archiving of data is widely spread, since the value of re-using data is not disputed. For instance, the Virtual Observatory is a new attempt at using archived observation data as a virtual observatory and thus as a low-cost possibility to conduct astronomical research. Basic needs for interchange call for a single standard format that is controlled by a standards body, independent of computer architecture, extensible, self-documenting, and remains fully backwards compatible. The FITS format was first introduced in 1979, and was formally endorsed by the International Astronomical Union, the supreme authority in astronomy, already in 1982. The IAU entertains a FITS working group that controls the standard.[21]

A key success factor for FITS is that the change procedures are intentionally complicated, involving regional committees and the FITS working group and requiring sound majorities. This helps sustain the key requirement that a FITS document will never become invalid.

The FITS structure, as created in 1979, reflects some of then current aspects of computer science. It is based on logical 2880 8-bit blocks to allow for different byte lengths. It consists of Header and Data Units (HDU). The header comprises an arbitrary number of fixed format 80 char card images. A data unit comprises multi-dimensional arrays in unsigned 8-bit integer, signed 16/32 integers, and 32/64 IEEE floating point numbers. To this, there are extension HDUs, such as image, ASCII table, and binary table extensions. There is a possibility to deal with redundancy through array and heap elements. A specific feature of astronomy is the multitude of available coordinate systems that need to be catered for.

FITS is universally used in astronomy. All major observatories provide observational data in FITS, as well as all data archives. Some of them also use FITS as their internal format. All software packages and data processing systems handle FITS; in fact, the community would accept no software tool that did not offer this possibility. These advantages must be weighted against some problem areas: the header has a very fixed format (because it is old) that makes support of hierarchical keywords and multi-value keywords impossible). Also, the format is very flexible, which raises the threat of bad design and requires very careful design of new file structures. That FITS only

[20] http://www.eso.org/.

[21] The International Astronomical Union homepage is at http://www.iau.org/, the homepage of the IAU FITS Working Group can be found at http://fits.gsfc.nasa.gov/iaufwg/. Some of the reference sources for FITS are the FITS Support Office at NASA/GSFC, http://fits.gsfc.nasa.gov/; the FITS archive at NRAO, http://www.cv.nrao.edu/fits/; and the Multimission Archive at Space Telescope's FITS page, http://archive.stsci.edu/fits/.

supports ASCII characters is no issue, since English has long become the universal working language of the astronomical community.

Further directions will include defining XML for small tables to increase flexibility and to incorporate a scheme for persistent unique identification, thus also enabling cross-references.

During discussion, it was pointed out that the digital preservation community might learn from this different domain. But, one has to consider that the astronomy community is likewise small, the needs of transferable and preservable standards are quite homogenous and widely recognized, and last but not leas, there may be not much economic interest behind format issues. Grosbøl added that the astronomy community also has a very open-minded access to data ownership, which sometimes causes problems with funding agencies.

Digitisation – The Only Viable Way to Preserve Audio Recordings in the Long Term

Introducing audio format considerations, **Dietrich Schüller**, director of the Austrian Academy of Science's Phonogrammarchiv (PhA), gave a short overview on the history of audio preservation. Audio (and video) materials are highly endangered by carrier deterioration and the obsolescence of format-specific hard- and software. Be it mechanical audio carriers, magnetic tape, or compact discs, a good number of carrier formats are currently at risk of being lost. For instance, professional video equipment is already dramatically endangered. The recognition that preserving carriers plus maintaining dedicated equipment of ever growing numbers of formats in playable condition is hopeless led the audio preservation community to a shift of paradigm in 1989/90, summarised in the motto "preserve the content, not the carrier." Audio (and video) preservation must therefore be based on subsequent digital, lossless copying of contents. Consequently, analogue holdings must be digitised. Radio sound archives took the lead (in Germany, first of all), to be followed by national and research sound archives in the later 1990s.

The 1990 vision were digital mass storage systems (DMSS), allowing for maximum automation of checking, regeneration, and migration, and, most importantly, offering new dimensions of access to collections. They include a combination of a hard-disk array and robotic tape store, are custom built, and very cost-intensive, in particular with regard to software. Small, scalable, modular solutions, so-called "Personal DMSS", would offer some relief.

The International Association of Sound and Audiovisual Archives (IASA) provides guidance through publications. "The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy", edited by the IASA Technical Committee (TC) 03 is available in version 2 (September 2001), and version 3 will be released in Summer 2004.[22] Also, Guidelines for the Production and Preservation of Digital Audio Objects are forthcoming in Summer 2004.

The IASA TC 03 has specified a number of requirements for digital preservation of audio recordings. These include:

- Optimal signal retrieval from analogue originals. Since digitisation is most likely a once-and-only event, the use of well-maintained, last generation equipment is indispensable.
- Unmodified transfer to new target format. While technicians usually try to make the best out of a record, this is not appropriate for preservation copies. The

[22] http://www.iasa-web.org/iasa0013.htm.

signal must be preserved free of alterations, de-noising, and similar improvements. The rationale behind this recommendation is that every modification will cause information loss. Additionally, it does not make sense to waste time and money on restoration at preservation time, given the fact that better tools and techniques will be available later.[23]

- Upcoming improvements of transfer technology recommend keeping the originals for possible later consultation. For instance, optical and contactless replay of mechanical carriers is being researched and will be available soon.
- Digital target formats. For digital audio, there is a de-facto standard, namely WAVE (.wav). This has been extended by the European Broadcasting Union to the Broadcast Wave Format (.bwf) to allow for limited metadata, and recommended by IASA. Another suitable, but not widely used de-facto standard is the Audio Interchange File Format for MacOS (AIFF).
- Digital target resolutions for analogue sources. Present practice employs a sampling rate of 48 kHz and a word length of 24 bit for radio sound archives, while for heritage and research archives 96 kHz/24 bit is appropriate. This results in storage requirements of one or two GB per hour, respectively. As the IASA TC 03 states, "the intended signal is only part of the sound document. […] Unintended and undesirable artefacts (noise, clicks, and distortions) are also part of the sound document, even if they have been subsequently added to the original signal by mishandling or poor storage. Both have to be transferred with utmost security." As for digital sources, they should be preserved as WAVE-files in their original resolution.
- Data reduction, better (but mistakenly) known as data compression, must not be employed for archiving. It is however useful and widely accepted for browsing purposes.
- Digital archival principles include: produce digital preservation copies free of uncorrectable errors or with lowest possible rate of correctable errors; produce and keep an error status report; check the error status at regular intervals, refresh before uncorrectable errors occur; migrate before systems/formats become obsolete; and produce and preserve a reasonable number of identical copies.
- Before DMSS become affordable, manual imitations of DMSS are an option. Dietrich strongly advised against the use of CD-R and DVD-R as storage media, since research has shown that their mean error rate already at recording time is prohibitively high. Rather, preservation institutions should chose computer back-up formats (such as DLT or LTO) immediately.

In summation, Dietrich Schüller underlined the predominant technical and strategic tasks ahead. The technical challenge is to feed estimated 100 million hours of analogue and digital audio holdings into a preservation environment. For classical transfer, one must count between three and six operator hours for the transfer of one hour of original material. This can be greatly reduced by "factory transfer", where one technician operates three to four semi-automatic workstations in parallel. However, a factory transfer is only possible for uniform source materials, and the investment costs for the workstations are currently around 60.000 EUR. The strategic challenge is to reach the (estimated) 80% of audiovisual materials that are outside of proper archival custody. To spot and preserve these collections of private owners, and small research and cultural heritage institutions, must be organised on a great scale.

[23] During discussion Dietrich pointed out that in the analogue part of digitisation noise reduction must be applied, but that any alteration must be avoided for the digital data.

In discussion, Dietrich Schüller specified that the currently excessive storage costs of around 10 EUR per GB and year are a consequence of past customers that were willing to pay any price. Storage software, that causes most of these costs, has to become more commercial and competitive. On the question whether BWF is actually an adequate preservation format, Dietrich underlined that not only its specification is open, but also there are around 30 Petabytes of BWF data expected for the foreseeable future, which alone will make it a prevalent standard. Although, it is not obvious that the additional internal metadata which distinguishes BWF from a pure Wave-File are really needed for preservation because they must as well be captured in a separate metadatabase in order to properly manage and retrieve files from the repository. Although redundancy is always prudent the BWF format requires in many cases conversion for access because most of the enduser software in the non-professional area is not able to read and render BWF-files.

Format and File Issues for Video Archiving

Experiences, best practices, and recommendations for audio formats were paralleled by a presentation on video formats, given by **Franz Pavuza**, also from the Phonogrammarchiv of the Austrian Academy of Science. He provided some general thoughts about video preservation, then went into the details of video preservation at the PhA.

Video footage is available at broadcasting companies, professional sources (such as commercial, industrial or academic institutions), and consumer sources (the latter becoming more and more important). Over fifty analogue formats and some 15 digital formats are currently in use or around. Two main archiving procedures are in use. Preservation of original media is facilitated because these are mostly tape, whose qualities and ageing process is well known. Transfer of content to a new carrier is the alternative.

Analog to digital conversion is very complicated for video because of the sheer mass of data. One hour of digital video can take up to 100 GB of storage space, and high transfer speeds are needed.

Three issues need to be treated: compression, storage media, and file formats.

Compression, although a popular, widely used, and cost effective strategy, must not be permitted for archiving analogue footage, as it entails a loss of information. However, there is nothing to be said against compression for browsing, streaming, and user copies.

As for media, only tapes are regarded as suitable for long-term archiving. Optical disks may be used for author copy and browsing, while magnetic disks are recommended to enable quick access.

Requirements for a video preservation format include platform and OS independence, extensive metadata support, multi-usability, expandability, not proprietary, standard, accepted by both manufacturers and users. Right now, the video preservation community is introducing a format that meets these requirements: the Material Exchange Format MXF, a file wrapper and packaging format capable of encapsulating video, audio, single pictures, etc., plus associated metadata. The MXF has been developed by the Pro-MPEG Forum and the G-FORS Group of the EU. It has received support from AAF and major vendors and institutions, and it has been standardised by SMPTE and EBU.

MXF has a wrapper for multi-media container. It follows a Key/Length/Value (KLV) format, where the key defines the data type, the length the size of the following block, and the value the actual content. This provides fast or partial access.

In the second part of his presentation, Pavuza introduced the approach the Phonogrammarchiv chose for its video footage. They use used, but well-serviced analogue players (e.g. from broadcasting institutions) and multi-format digital players, together with a PC based capture station plus additional dedicated hardware. For browsing, they use VHS tapes and DVD (eventually to be replaced by a file server), where they propose selected files in 10MB/s quality, and all files in a standard, 1MB/s quality. For archiving, LTO tapes have been chosen. These have been developed as computer backup tapes. They are highly reliable and designed for self-checking, widely accepted, and costs are rapidly decreasing, having reached some 50 EUR per 100 GB. Currently, the first generation is being used, offering 100 GB with 15 MB/s access speed. A second generation is being introduced, while the third and fourth generation (800 MB) are expected within three years. The generations are to be fully compatible.

Speaking about software, Pavuza acknowledged that there is no dedicated video archiving software, which leads the PhA to use editing software.

An alternative solution would be to “rent an archiving system”. This could automatically check, clean, record, and transfer video files at reasonable costs, if the number of tapes is high and there are only a few standard formats in use. Main problems include the analogue players, the lack of archiving software, the workflow and time requirements, quality control, and high definition.

Conclusion

The workshop showed that the situation concerning file formats for preservation differs very much depending on the type of digital objects and the area and sector where they are used. While it has been apparent that is easier to maintain and develop a format for preservation in a close-knit and small community where there is a strong interest in a common preservation format, other domains like the ordinary office automation (word-processing, spreadsheets, etc.) find it far more difficult to discover an appropriate file format for preservation. Where market development and competition are very strong, requirements for preservation are only addressed if they help to increase market share, and that is, unfortunately, not often the case.

Preservation formats will not last forever. They certainly will evolve as digital objects will evolve. Text documents which are today in most of the cases reducible to two-dimensional images will in future no more be convertible in that way without unacceptable loss of information. The choice of preservation formats has therefore a lot to do with medium and long-term IT-planning. On the other hand it makes no sense to use a specific file format for current business if one does not know how it can be appropriately preserved, i.e. converted in a stable preservation format.

Preservation formats therefore need to be maintained, format developments need to be carefully watched in order to keep always a migration path open to new formats when old formats will become obsolete.

List of Participants

Caroline Arms	Library of Congress	USA
Stefan Arnold	Fabasoft AG	Austria
Andreas Aschenbrenner	ERPANET	Netherlands
Christoph Bauer	ORF – Österreichischer Rundfunk	Austria
Chris Bellekom	National Library of the Netherlands	Netherlands
Olaf Brandt	Staats- und Universitätsbibliothek Göttingen	Germany
Adrian Brown	National Archives	UK
Jane Brown	National Archives of Scotland	UK
Georg Buechler	ERPANET	Switzerland
Niklaus Buetikofer	ERPANET	Switzerland
Edgar Büttner	Bundesarchiv	Germany
Cinzia Cappiello	Politecnico di Milano	Italy
Lars Clausen	State and University Library	Denmark
Remy Cristini	Sinological Institute, Leiden University	Netherlands
Claus Dam	Danish Heritage Agency	Denmark
Richard Davis	University of London Computer Centre	UK
Michael Day	UKOLN, University of Bath	UK
Stuart Dempster	JISC	UK
Aranea Dijkmans	Municipal Archives Amsterdam	Netherlands
Jim Downing	Cambridge University Computing Service	UK
Jayne Dunn	Natural History Museum London	UK
Katharina Ernst	Stadtarchiv Stuttgart	Germany
Miguel Ferreira	University of Minho	Portugal
Lars Gaustad	National Library of Norway	Norway
Guido Goedemé	Koninklijke Bibliotheek van België	Belgium
Preben Grosbøl	European Southern Observatory	Germany
Mary-Ann Grosset	OECD	France
Maria Guercio	ERPANET	Italy
Rachel Heery	UKOLN	UK
Helen Hockx-Yu	JISC	UK
Hans Hofman	ERPANET	Netherlands
Thomas Huemer	Austrian Academy of Sciences	Austria
Vanya Ilieva	New Bulgarian University	Bulgaria
Hamish James	Arts and Humanities Data Service	UK
Dirk Janssens	OECD	France
Jon Juliusson	Swedish Tax Agency	Sweden
Thomas Just	Österreichisches Staatsarchiv	Austria
Max Kaiser	Austrian National Library	Austria
Leopold Kammerhofer	International Atomic Energy Agency	Austria
Bettina Kann	Austrian National Library	Austria
Christian Keitel	Staatsarchiv Ludwigsburg	Germany
Uwe Klosa	Uppsala University Library	Sweden
Ellen Konstad	Det Norske Veritas	Norway
Wolf-Dieter Lang	Austrian National Library	Austria
Tue Hejlskov Larsen	The Royal Library	Denmark

Hanno Lecher	Sinological Institute, Leiden University	Netherlands
Karl-Ernst Lupprian	Generaldirektion der Staatlichen Archive Bayerns	Germany
Despoina Manolopoulou	Aristotelian University of Thessaloniki	Greece
Amy Marshall	Art Gallery Ontario	Canada
Adelheid Mayer	University of Vienna	Austria
John Mc Donough	National Archives	Ireland
Paul Meinl	Factline Webservices GmbH	Austria
Frank Moehle	Swiss Federal Archives	Switzerland
K. Helen Moeller	Det Norske Veritas	Norway
Christa Müller	Austrian National Library	Austria
Eva Müller	Uppsala University Library	Sweden
Hannes Obermair	Stadtarchiv Bozen	Italy
John Mark Ockerbloom	University of Pennsylvania	USA
Osmo Palonen	Mikkeli Polytechnic	Finland
Franz Pavuza	Phonogrammarchiv	Austria
Christine Pétillat	French National Archives	France
Evangelia Psourouka	Alfa Asfalistiki	Greece
Johanna Rachinger	Austrian National Library	Austria
Andreas Rauber	Technical University of Vienna	Austria
Carl Rauch	Technical University of Vienna	Austria
Heinz Reim	Austrian National Library	Austria
Mark Richard	Swiss National Library	Switzerland
Peter Rinzner	Magistrat Wien	Austria
Bill Roberts	Tessella Support Services	Netherlands
Seamus Ross	ERPANET	UK
Dietrich Schüller	Phonogrammarchiv	Austria
Anette Seiler	Hochschulbibliothekszentrum Köln	Germany
Jordi Serra Serra	Arxiu Central, DURSI, Barcelona	Spain
Robert Sharpe	Tessella Support Services	UK
Martin Slabbertje	Utrecht University Library	Netherlands
Jacqueline Slats	Nationaal Archief	Netherlands
Josef Steiner	Austrian National Library	Austria
Tobias Steinke	Die Deutsche Bibliothek	Germany
Martin Stürzlinger	Wiener Stadt- und Landesarchiv	Austria
Susan Sullivan	NARA	USA
Jean-Pierre Teil	French National Archives	France
Matthias Töwe	Consortium of Swiss Academic Libraries	Switzerland
René Van Horik	NIWI-KNAW	Netherlands
Willy Vanderpijpen	Royal Library of Belgium	Belgium
Bert Wendland	Humboldt University Berlin	Germany
Martin Wynne	Oxford University	UK
Thomas Zäschke	State and University Library	Denmark