



**Project no.507618**

**DELOS**

**A Network of Excellence on Digital Libraries**

**Instrument: Network of Excellence**

**Thematic Priority: IST-2002-2.3.1.12  
Technology-enhanced Learning and Access to Cultural Heritage**

**<DELIVERABLE REFERENCE NUMBER: WP6, D6.3.1  
File formats typology  
and  
registries for digital preservation>**

Due date of deliverable: 31 December 2004  
Actual submission date: 15 December 2004

Start Date of Project: 01 January 2004  
Duration: 48 Months

Organisation Name of Lead Contractor for this Deliverable  
UNIVERSITA' DEGLI STUDI DI URBINO –ISTITUTO DI STUDI PER LA  
TUTELA DEI BENI ARCHIVISTICI E LIBRARI (ISTBAL)

**Revision [2.5]**

Project co-funded by the European Commission within the Sixth Framework  
Programme (2002-2006)

---

Dissemination Level: PU

**UNIVERSITA' DEGLI STUDI DI URBINO –  
ISTITUTO DI STUDI PER LA TUTELA DEI BENI ARCHIVISTICI E  
LIBRARI (ISTBAL)**

**FILE FORMATS TYPOLOGY AND REGISTRIES FOR DIGITAL  
PRESERVATION**

**DELOS REPORT**

**Professor Maria Guercio**

**&**

**Cinzia Cappiello**

**Abstract:** With the development of technology, file formats are increasing in both number and complexity. This is a critical issue for digital preservation. The aim of digital preservation is ensuring that records are filed and made accessible throughout time, but as a result of progress in software and technology old formats soon become unreadable and unusable. Different research initiatives are focusing on this issue, trying to define preservation-friendly standard formats, as well as strategies for records to be made available over time and for their presentation and content to be preserved. This report offers an overview of file formats presented in the relevant literature and of their characteristics. The fundamental requirements for file formats in a digital library environment are identified and discussed. In addition, the most important digital preservation strategies for managing file formats are presented. To complete the analysis of the relevant literature, the most significant projects dealing with file formats management are then illustrated.

*Task Leader and Contact: Maria Guercio, UNIURB (30)*

*Participants: UNIURB (30), UG (9), OEAW (37), UKOLN (4), TUW (29)*

---

## **Preface – Scope of the document**

In the literature, file formats are largely studied in different fields and for different purposes. The number of file formats keeps increasing, as new and different software programs for different operating systems are developed and released. Each software utility can handle a given set of file formats. And within each set of file formats, we may distinguish the main file format - the format in which the information is commonly saved - and a definite number of compatible file formats that the program can manage.

It is easier to exchange information when file formats are portable. The portability of a file format can be measured by two different indicators: the number of software programs that are able to handle it and - whenever a file format is associated to a specific software program - the capability of that program to run indistinctly on different platforms. Today, there are many works to be found within the file formats literature focusing on this topic and evaluating the various degrees of file formats portability. Whereas other works have focused on classification systems and on the building up of proper file formats registries. Such registries should provide something similar to a file formats “census”, for users to be able to associate to any given file extension the correct file format and the software programs that can handle it.

This report wishes to review the file format literature under a digital preservation perspective. Digital preservation aims at ensuring the accessibility of information throughout time. The first objective of this report is exploring the relation between file formats and digital preservation, highlighting which characteristics make a file format a preservation-friendly format. We believe this could support research on file format registries, and contribute in understanding which could be the new mechanisms for using registries in the digital preservation domain. Secondly, we shall discuss the most important digital preservation strategies: this report wishes to offer a possible taxonomy, based on the way in which information and its related file formats are handled. Finally, by studying the main characteristics of file formats and digital preservation strategies, this report also wishes to define a reference framework of typologies and of specific rules for specific cases, which may be used to identify

suitable solutions capable of taking the specific preservation context and users' requirements fully into account.

The literature review has been difficult, considering the large variety of fields in which this topic is studied and the huge number of existing file formats in use for different typologies of data. The relevant literature offers an analysis of the different file formats related to specific types of data, i.e. images, defining indicators for their evaluation in terms of specific aspects such as that of loss of details when compressing files. There are only a few works that have carried out an extensive overview of file formats in a specific domain, such as that of digital preservation.

## **1. General Introduction**

Digital information is available in a large number of standards and proprietary formats. Each format evolves continuously, becoming more complex and powerful in terms of its functionalities. A strategy for handling file formats is an essential part of any effort of long-term preservation of digital objects. Evolution in software and hardware technologies is an obstacle to digital preservation, as old formats and documents become unreadable and unusable. It is thus necessary to take into consideration and study methods and approaches aimed at ensuring that digital objects will still be readable and understandable after a given length of time.

Many projects are focusing on these issues, proposing different solutions for digital preservation. The main difficulties are related to the increasingly complex nature of stored information and to the subsequent mushrooming of file formats. Today, stored records would not only contain text but also formulae, charts and multimedia files, like images, sounds and videos. As a result of such a heterogeneous nature of storable information, a high number of file formats are now spread, and many of them often need a specific software for being used and are scarcely portable. Consequently, at the current pace of evolution in both software and hardware, many file formats are likely to quickly become obsolete and unusable.

At the moment, there are a lot of initiatives working on the definition of a general open format for documents. In this sense, interoperability comes as a problem strongly related to preservation – and, actually, research on this type of format originally started because of the need to improve the degrees of interoperability of the information's content. The problem of preservation can be also studied from the

interoperability perspective, since the interoperability problem focuses on *wideness of use*, while the preservation problem focuses on *wideness over time* - that is to say, *digital longevity*. If we wish to guarantee digital longevity, we must look at how preservation in the digital world differs from what we are accustomed to in the real world [22]. According to H. Besser, there are five key factors giving rise to digital longevity problems: the Viewing Problem, the Scrambling Problem, the Interrelation Problem, the Custodial Problem, and the Translation Problem [22]. The *Viewing Problem* regards the need for an infrastructure and a common knowledge allowing us to view digital information that was created in the past. The *Scrambling Problem* looks at initiatives which might have been taken in the past to solve short-term problems related to digital information, but which ended up creating difficulties in the long-term. A clear example of this kind of initiatives is the adoption of file formats that use compression, i.e. JPEG format: with this kind of formats, one gains in terms of disk space, but loses information. The *Inter-relation Problem* looks at information which is related to other information, such as in the World Wide Web context. Interconnections between documents are difficult to manage, and it often happens that when a record changes in location or content, nobody updates its connections and relations to other documents, and information gets disconnected and lost. The *Custodial Problem* regards the identification of responsibility in the preservation domain. Finally, the *Translation Problem* deals with losses in information during the translation from a specific file format to another.

All these aspects are studied by various research schools active in the digital preservation domain. From the specific perspective of file formats, digital preservation theories are often based on simple assumptions [25], such as: XML is the most suitable format for storage, metadata are essential, standards will be the solution to all our problems, data storage media are robust, registries are essential, etc. This report aims to evaluate these and other similar assumptions which can be found in the relevant literature, and it will try to cover all major contributions on the issue of file formats and digital preservation.

This paper is structured as follows: Section 2 describes a categorization of file formats, which is useful to discuss the relevant aspects of preservation and conversion. Section 3 introduces a series of aspects of files which may be preserved through various techniques, and makes some considerations on the possible criteria

for assessing the middle-term usability of a file format. Finally, section 4 introduces the most relevant preservation strategies.

## **2. File Formats**

### **2.1 Introduction**

A file is a sequence of bits, whereas a file format is a format for encoding information on a file [26]. Since a disk drive, or any other computer storage system, can only store bits, the computer must have a way to convert information in zeros and ones and vice-versa. There are different kinds of formats for different kinds of information. Each different type of file has a different file format. The first thing a file format specifies is whether the file is binary or ASCII, and second is how information is organized. To be more precise, file formats are composed of a file header (metadata) and the data themselves. The file format is one of the keys to access the information object. File formats must be analysed together with the software required to create, manipulate or convert the format, and with the hardware on which the operating system works. In order to define a file format, we have to specify the degree of granularity which we wish to consider in the file format specifications: file formats are often encapsulated in one another, as in the case of an image included in a PDF document. In order to give any formal definition of file format, the level at which file formats are analysed must be specified.

### **2.2 Categories of file formats described in literature, and remarks on their dynamicity, volatility and inconsistency.**

Looking at the current situation in the file formats domain, the first thing one notices is unquestionably the high number of formats available. The File Extension collection [30] asserts they have indexed over 15,000 file name extensions, and more advanced indexes such as [31][32][33] already contain something like a thousand descriptions of file formats. Considering file formats listed in [33], the category that contains the largest number of file formats (42%) is the one related to graphics file formats. 30% of file formats are related to programs and utilities. Lower percentages are related to sound formats (8%), text formats (8%), video formats (3%), and database formats (9%).

The number of file formats is incredibly high even if we limit our attention to specific families of file formats (i.e. Word). Programs don't know how to manage all the import-export filter operations to all available formats. This would be rather expensive in terms of acquisition and maintenance procedures, with a maximum risk of incompatibility and error. In addition, we must consider that there is a close, one-to-one dependence between each format and its parent tool [1].

Moreover, we must remember that one of the main advantages of electronic information is the fact that it can be easily stored, accessed, and re-used. The choice of the file format has critical consequences on the possibility to archive and preserve information. It is generally recommended to use as many standards as possible for electronic records, and to avoid archiving data in application or proprietary formats [15], since these are always related to a particular provider, computer application or even version. Each file format presented in this section will be discussed also in terms of preservation issues.

The purpose of this chapter is to provide a survey on file formats today available, highlighting which are the most important so as to define the current context.

We may call the most used type of file a '*Document-like*' file <sup>1</sup>. (Remember that in this report only *electronic* documents are considered.) In this perspective, the document is generally a file associated to a specific topic and composed of different objects such as text, graphic objects, and multimedia objects. The document is associated to a precise file format that encapsulates other formats.

With reference to different collections of file formats, one can identify more than one series of categories. Most of them are not directly related to preservation aspects, but are aimed at facilitating things for users.

The simplest categorization is in the *File Extension* collection [30], which only operates a distinction between graphical and non-graphical formats. The *Wotsit's Format* [31], *My File Formats* collection [32] and the *File Format Encyclopaedia* [33] are instead based on a rich set of categories, classifying file formats as follows:

- Graphics file (BPM, AVI, GIF, JPG...)
- 3DGraphics file (3DX, M3D, SDML, VRML....)

---

<sup>1</sup> A document-like object is a digital data unit that is comparable to a paper document. The term designates a relatively simple stable resource, and would not cover, for example multimedia artefacts or interactive services [Definition provided by Open Archive Forum].

- Movies, Animation (MPEG, QUICKTIME,....)
- Archive files (ARC, ARJ, ZIP....)
- Binaries (EXE, LIB,...)
- Spreadsheet/databases (MDB,XLS...)
- Financial/stocks (Metastock, Quattro...)
- Font files (FNT, TEX...)
- Game files (DAP,DEM...)
- Text files/document (DOC, EPS, PDF...)
- Internet related (CooKIES, DBX, CSS...)
- Sound and music (MP3, DIGI, WAV....)
- Windows files
- GIS files

The official categorization of file formats is the MIME type, provided by IANA [34].

They define the following main categories:

- application,
- audio,
- image,
- message,
- model,
- multipart,
- text,
- video.

According to R.L. Clausen R.L [29], any of these categorizations overlap at least once: for example, when using the MIME standard the RTF format exists in both the “application” and the “text” category. For this reason, in [29] the author identifies a categorization where one doesn’t have to define the “application” category. File formats are defined according to the following classes:

- Document-like (PDF, DOC, PS, DVI, HTML...)
- Image formats (GIF, PNG, JPG)
- Sound formats (MP3, OGG,....)
- Movie formats (MPG, AVI...)
- Data formats (more or less raw data from experiments)

- Structured graphics format (XSL,...)
- Databases (DBF, DDF,...)
- Collection (tar, zip, ....)
- Configuration and metadata (CSS,...)
- Program-supporting formats (TTF, game saves,...)
- Program file formats (Javascript, Java, SWF ....)

In the following table, we compare the classifications introduced in this section, so as to better understand the way categories are related to each other.

<b>Statsbibliotek et (Clausen 2004)</b>	<b>MIME</b>	<b>Wotsit's Format</b>	<b>MyFileFormats</b>	<b>File Format Encyclopedia</b>
Image/structured graphic	Image	3D Graphics Files	3D Graphics Files	3D Graphics Files
		Graphics Files	Graphics Files	Graphics Files
Movie	Video	Movies/Animation	Animation and Movies	Animation
Collections		Archive Files	Archive Files	Archive
Program files	Application	Binaries	Binary files	Binary Emulator
		Comms Format	Comm files	-
Spreadsheets		Spreadsheets/Databases	Spreadsheet files	-
Databases			Database files	Text
Document-like	Text	Text, Files/Documents, Printer formats	Documents and text files	Text
Data Formats	Message	Financial/stocks	-	-
		GIS Formats	GIS files	
Program supporting	Application	Font files	Font files	Font
		Games files	Games files	Games
-		Hardware formats	Hardware formats	-
-		Internet related	Internet files	-
-		Miscellaneous	Miscellaneous	-
Sound	Sound	Sound and music	Music and sound files	Sound/modules
-		Windows Files	Windows files	Windows

The main differences between the classifications are justified by the different purposes with which they have been built. Neither the *Wotsit's Format* collection [31], the *My File Formats* collection [32] nor the *File Format Encyclopaedia* [33] were conceived for preservation needs, but for facilitating the lives of users exploring all the existing formats. Details in classification are in fact needed to simplify searches and increase the usability of the Web Site in which the file formats are listed. As discussed above, in these classification systems categories often overlap, and a single file format can belong to more than one category. On the contrary, the classification proposed by Clausen has a greater focus on preservation and tries to avoid the overlaps and confusion of the above mentioned categorizations. Yet even here categories would sometimes overlap. The overlap is actually intrinsic in the definition of the *document-like* category. Indeed, it includes all the file formats able to encapsulate different types of objects in a unique source. Just think of the most famous document-like formats, such as Microsoft word format or PDF format: these formats can handle, within the same file, different types of objects. The most common combination is, for example, text and images. Another categorization which takes the preservation issue into account is the one presented in a work on file formats prepared by the European Community [1], where only the four categories of sound, image, movie and document-like file format are defined. In the preservation perspective, this kind of classification is actually very correct, as these categories are sufficient for exploring all the major aspects related to the preservation issue. For what concerns this report, we have tried to distinguish between file formats that can manage heterogeneous objects and formats that are uniquely related to a precise type of data. For this reason, file formats are distinguished between:

- *Compound file formats*, which can manage the aggregation of basic file formats. This category contains *document-like and scientific file formats*.
- *Basic file formats*, which can manage only a specific type of object. They can be related to *text, sound, movie, image, or database formats*.

We believe the formats considered in this classification have been deeply examined and are critical for the preservation function. In all the preservation strategies and projects which we analysed, the main open research questions are related to these formats. The preservation of program files is usually made through emulation or

technology conversion. The file format in this case is always the original and the focus is on preserving the technology and software used for emulation.

### **2.2.1 Compound file formats**

The compound file formats are able to encapsulate different objects that are identified and stored using basic file formats. The aggregation of various basic file formats is managed in order to provide different types of information about a specific topic that is manageable and accessible through a unique source. Document-like formats and scientific formats are contained in this category and discussed in the following sections.

#### **2.2.1.1 Document-like formats**

The main goal of document-like formats is handling information on a specific topic in a digital format, through the simulation and use of editing and representation techniques which are typical of paper documentation.

Considering the duration of a digital document, we can distinguish between formats related to:

- *Documents created on the fly*: documents in which the content is temporary. Sometimes content in this type of document lasts for only one working session, and is usually determined by a certain combination of factors such as the specific temporal sections or a set of user preferences. These documents are generally produced by tools that work on the appearance of the document given by the representation of the text. The presentation layer and the content source are often kept separate, in order to provide a high flexibility in the creation and composition of the document. Typical examples of these tools are Web-oriented editors.
- *Documents created permanently*: the content is permanently stored inside these documents. Both the structure and content are usually defined at the moment the document is created, by using tools that work on the abstract internal representation of the document. Typical examples of these formats are the PDF format or the Microsoft Word format.

Representing text is one of the primary focus of a document. However, most word processing tools would not focus on the printed image of the document, but rather on their internal representation. These tools manage the structure of documents, they use characters and codes to distinguish words, paragraphs, chapters, pages, etc., and they assign markers and define the styles of these objects, finally producing in real-time an on-screen image from the internal representation. Working with text and attributes is easier than working with a text document at the level of printed image, and changes to documents are always made at this abstract level. Only a small set of word processing tools works with the “What You Can See is What You Get” criteria, producing an on-screen image which is very similar to the one that will be produced by the algorithms that put the ink on the printout.

As any other program, a word processing program works with many variables and data structures, within defined and dynamically allocated storage areas. The status of the programs and text, along with its associated properties, are saved in the computer’s main memory. All these elements belong to a dimension of internal representation, and are needed to reconstruct the internal representation from the file itself. The easiest way to store the internal representation would be a complete storage dump. In this case, the stored data is only significant to the program that wrote it, and results in a proprietary document format. Some proprietary document formats have implemented a minimum level of functionalities which guarantee that the document can be opened on different computers. However, they remained incompatible between different versions and platforms.

In order to solve incompatibility problems, word processors have introduced import and export filters, which allow to import and export documents to other formats. Usually, what the word processors try to guarantee is the backwards compatibility, even if the results of conversion in older formats often determine the loss of information, and consequently the degradation of the format. This does facilitate the exchanging of documents in an increasingly networked and collaborative business environment. Therefore, it is at this point extremely important to standardise the document format by publishing its specifications and making them available. Whenever a file is saved, the internal representation of the text document is converted to its standard format. Inversely, when the file is read by another tool, the format is abstracted and converted into its internal representation.

A part from text, electronic records can contain many types of components. Users may integrate various types of graphics, diagrams, multimedia files or add spreadsheets. In addition, data which are related to the main document but are not part of the content must also be managed. These data could include, for example, information about the configuration of the program or the format templates.

Moreover, one might also have a reason for needing to store additional information. This information comes in the form of metadata, and may be about the record itself, the manipulations it underwent or its users. As already mentioned, metadata are particularly relevant in the field of file preservation, since the overall information on those document has to be preserved. For being adequate for preservation, any format must be capable of managing metadata efficiently.

In the following, the main document formats are discussed.

#### **2.2.1.1.1 XML (eXtensible Mark-up Language)**

The XML (eXtensible Mark-up Language) was developed in 1996 by a XML Working Group created within the World Wide Web Consortium (W3C). XML uses a few associated standards, namely Unicode and ISO/IEC 10646 for characters, Internet RFC1766 for language identification tags, ISO 639 for language name codes, and ISO 3166 for country name codes. XML is used for describing marked-up electronic texts. XML is a metalanguage, that is, a means of formally describing a mark-up language. A mark-up language specifies what mark-up is required and allowed, and defines the mark-up and how it can be distinguished from text.

XML has three main characteristics [1]:

a) Emphasis on descriptive rather than procedural marking-up. The idea is that XML is trying to apply markups that with the name suggest and describe the text to which they are applied. XML provides a nice flexible internationalized way to label the elements of a data structure and ship them around with the labels attached. In descriptive mark-up systems related to the single entity, mark-up codes or tags are basically only used to identify and categorize parts of a document. They generally define what the element content is, rather than how it is processed. In the XML world, the instructions needed to process a document for some particular purpose are kept separate from the descriptive mark-up. They are collected outside the document in

separate procedures or programs and are usually expressed in a distinct document called stylesheet, which doesn't simply contain information on the visual appearance of the document. The same document can be rendered or processed differently on multiple channels and for multiple users profiles. The document can be archived with no risk of stumbling in machine dependent processing instructions.

b) The document type concept. The type of a document is defined as a DTD or a XML Scheme. If the document's type is already familiar, a special purpose program (called Parser) can check the conformation of the document in that specific format.

c) Independence from hardware or software system. One of the main goals of XML is to ensure that records can move from a hardware or software environment to another without losing information. This is possible because all records use the same standard, which is implemented by a universal character set maintained by an industry group called the Unicode Consortium.

The main characteristics that make XML different from HTML are:

- Extensibility: XML does not contain fixed tags.
- Flexibility: XML allows user to define an infinite number of DTD documents.
- Content-focused: XML focuses on what data is, not on how it is presented as a web page.
- Modularity: XML provides mechanisms to separate data from processed content.

In theory, when XML documents are created in a correct way, they should in theory be platform-neutral. Yet, vendors often try to maintain their identity, linking applications to the platform suites, with the result that documents based on different XML based formats are often not compatible, and sometimes the conversion between two different formats can be even impossible.

Nowadays, XML is becoming the standard for data exchange between information systems. But it is important to underline that XML is not a format in its own right, but rather something which enables the definition of formats. The big advantages of XML are its transparency and interoperability. In the past, precursors to XML file formats were limited in the data import and export operations. Indeed, one of the critical issue when an upgrade or a migration occurs, is whether the new arrangement will import the old files. For this reason, vendors are producing proprietary file formats to force user loyalty and at the same time, doing flexible tools to accept the files of competitors. But this obstacles the users to have the total control on their own data.

XML is proposed as solution because it is plain text, and consequently highly portable, and it is offered with tools that allow the conversion between different XML formats.

There are two relevant formats based on XML, the Microsoft XML Reference Schema and OpenOffice.org. In 2003, Microsoft announced that the company was releasing the XML Reference Schema for its office suite. Except for some technical details, the XML formats seem to be completely documented and the license to use them should be free, even if some constraints could not be avoided.

The OpenOffice.org Project produced both a mature, complete, open source, front office application suite derived from StarOffice and it is planned to switch the storage format to XML in the near future. The full documentation is freely available from the OpenOffice.org open source community [45]. OpenOffice includes applications that manage and produce formats suitable for office documents such as a word processor, spreadsheet, a presentation tool, and a graphics/diagramming tool. It is compatible with XML specification and it keeps the document's content and layout information separate such that they can be processed independently of each other..

#### **2.2.1.1.2 Microsoft DOC formats**

The DOC formats are one of the most widespread formats within the text format set. The document format of Word follows the internal program representation and is composed of multiple *binary* data streams:

- A summary information stream, containing summary and metadata.
- A main stream, containing the text and the formatting information.
- A table stream, containing coded references in table format between all data structures of data stream and summary information stream.
- A data stream, containing graphics and any embedded object in its native format.

#### **2.2.1.1.3 RTF**

RTF belongs to Microsoft, which created it as a uniform text exchange format where graphic integration is enabled. It is a method of encoding formatted text and graphics

for easy transfer between applications. Indeed, RTF provides a format that can be used with different output devices, operating environments, and operating systems. The software that turns a formatted file into a RTF file is called writer. An RTF writer separates the application control information from the actual text and writes a new file in which there is the text and the RTF groups associated with that text.

An RTF file consists of unformatted text, control words, control symbols, and groups. RTF is also suitable for the publication on the documents over the Web since it cannot save macros which make it invulnerable for macro viruses. RTF does not support password protection or encryption and it suffers the same problems related to backward compatibility as Word.

#### **2.2.1.1.4 PostScript**

PostScript is a programming language thought for printing graphics and text. It is a page description language. The main purpose of this language was to provide a language to describe images in a device independent way. Indeed, the image description is made without reference to any specific device feature, so that the same description could be used on any PostScript printer without modification. PostScript is a non re-writeable format.

#### **2.2.1.1.5 HTML**

HTML is an international language that allows authors to publish online structured documents with headings, text, tables, lists, photos. It is also possible to link the information in order to build a hypertext and to design forms for conducting complex transactions, such as searching for information, making reservations, with remote services. HTML is designed to be used on a multi-device platform: it has been developed with the idea that all devices should be able to use information on the Web.

#### **2.2.1.1.6 AdobePDF**

Portable Document Format (PDF) is a file format developed by Adobe Systems for representing records independently from the original application, software, hardware, and operating system used to create those very same records. A PDF file can describe records containing any combination of text, graphics, and images in a device-independent and resolution-independent format.

PDF results from the combination of three technologies:

- A form of PostScript for generating layout and graphics,
- A font-embedding/replacement system to allow fonts to travel with the documents,
- A structured storage system to bundle these elements into a single file, with data compression where appropriate.

PDF is a subset of PostScript language elements that define the graphics, and it only requires a very simple interpreter. The PDF format offers a great deal of advantages. First of all, PDF files are smaller in size than PostScript generated files. Secondly, thanks to advanced mechanisms you can visualise a PDF file even when the end-user hasn't installed the needed font on its computer. Hence, portability and independence of platforms are always guaranteed. Just as in the case of the PostScript format, PDF remains an end-form format and wasn't conceived for re-writing documents.

Some great results have been achieved with the PDF format world in terms of document's layout maintenance, but at the price of sacrificing the possibility to modify documents. The on-screen presentation and print outputs of PDF files reach the industry's highest level of fidelity, yet they are not totally perfect. The PDF encoding utility configuration can reduce the size of the generated PDF file, but also influences the quality and portability of the document: compression may degrade quality of the embedded images, since the applied algorithms sometimes lose information. Note that PDF is not adequate for being used as an archival format [20], and actually a PDF/A standard version has been specifically introduced for archiving purposes. It contains mandatory, prohibited or recommended components expressly created in order to address preservation needs. The PDF/A format is introduced in Section 2.4.4.

#### **2.2.1.2 Database file formats**

A database can be defined as a collection of data created for a specific purpose and accessible for one or more computer systems. There are different types of databases:

- *Relational database*: data are organized in tables linked by relations. Oracle Databases, Microsoft Access are the most popular systems in this area.
- *Hierarchical database*: data are organized in the form of a tree structure. Adabas is an example of hierarchical database management system.

- *Native XML database*: stores XML in "native" form, generally as some variant of the Document Object Model mapped to an underlying data store. Content management systems are often built on top of native XML databases.
- *Object database*: data are stored as objects and can be interpreted only using the methods specified by its class. The relationship between similar objects is preserved (inheritance) as are references between objects.
- *Network database*: a kind of database management system in which each record type can have multiple owners. This contrasts with a hierarchical database (one owner) or relational database (no explicit owner).

The critical issue in database preservation is that each database system is unique, and it is difficult to use the migration technique to convert the original database in a preservation friendly format. The big problem is that the application that manipulates and accesses the final database has to be custom developed in each case. It is also difficult to assure the reliability of the converted database since it is complicated to maintain the original structure in terms of links, joins, indexes and user issues.

### **2.2.1.3 Scientific file formats**

Scientific data management is critical in many areas of science and engineering. Today's computational and scientific techniques are capable of generating large amounts of highly complex data which must be effectively preserved and managed for future analysis and distribution. Key characteristics of scientific data are the highly dynamic nature of the data and data schema and also the dynamic nature of the formats in which the data are disseminated. Furthermore, scientific data themselves are complex since they include large, multidimensional arrays of numbers and the details are essential, since every number matters.

Considering the preservation, it is worth to discuss two relevant initiatives in the scientific data field: the XML introduction in the scientific data representation and the use Common biometric Exchange File Format (CBEFF) in the biometrical data management.

#### **2.2.1.3.1 XML for scientific data**

Scientific data are usually stored and transferred using a variety of data formats. In the past the most spread approach in dealing with scientific data was translating them in binary format. In recent years XML has become an important and popular language for exchanging digital information and it is largely used as lingua franca in the scientific data management [51]. The unique big disadvantage of XML in the scientific data management is its weakness in the representation of large arrays of numbers and structures other than a tree.

Translators that create a standard XML file from binary data, and create a binary file from appropriate XML have been largely developed. In essence, these translators are programs that read all or selected objects in the binary, and create an XML file which represents the same object according to some DTD or Schema. In a parallel way, considering a XML file these translators create the appropriate binary object.

XML is suitable for scientific data since it is flexible and it can model numeric concepts such as vectors and arrays in several ways because it is possible to “mark up” the same document using many different schemas. Further XML enables interoperability among different formats, indeed XML is well suited for heterogeneous systems with multiple users and multiple purposes. The universality and portability of XML is essential for sharing data across space (geographic distribution), time (archiving), and conceptual domains (different users, communities, and uses). Finally, an XML document can be used to deliver a description of the data without the data itself, since XML is able to link external data of different nature including binary files or objects in a binary file. And when there are large amounts of data, an XML description is preferable to the data itself.

#### **2.2.1.3.2 Common Biometric Exchange File Format (CBEFF)**

Common Biometric Exchange File Format is a format that enables interoperability of biometric-based application programs and systems from different vendors.

Indeed, the main goals of this biometric file format is handling different types, versions, and technologies of biometric data in a common way, and providing forward compatibility for technology improvements.

The structure of a common biometric file format is composed of a header that contains information such as file length and biometric types, and of a block of data (data structure) in unspecified format. In the CBEFF the header is called *SBH* (Standard Biometric Header) and it specifically contains general information about the ownership and the nature of biometric data that are stored in the file. The data structure is called *BSMB* (Biometric Specific Memory Block) and it is simply a block of memory in which data are represented in a format defined by the owner.

## **2.2.2 Basic file formats**

Basic file formats are created to manage a specific type of digital objects. In this report text, graphics, video and audio file formats are studied.

### **2.2.2.1. Graphics File Formats**

Graphics files may be defined as files that can store any type of persistent graphics data (as opposed, for example, to text, spreadsheet, or numerical data), and which are thought for possible rendering and display. There are three general classes of graphic file formats: bitmap or raster, vector, or hybrid files.

The *bitmap (raster)* files organize images as rows and columns of dots. The value of each dot is stored in one or more bits of data. The number of bits used to represent a dot depends on the colours and their specific tone. The resolution expresses the density of the dots and determines the goodness of the image. Bitmap files have the disadvantage to be not very flexible, indeed they are resolution and device dependent.

The *Vector files* describe images as mathematical formulas that define all the shapes involved in the image. This kind of graphics are more flexible than bitmap graphics, they are not resolution or device dependent and can be easily resized and manipulated on any device, and any resolution.

The *Hybrid files* combine vector and raster data in a unique image.

In the following sections the most common types of graphic file formats are briefly described.

### **2.2.2.1.1 Bitmap File Formats**

#### **2.2.2.1.1.1 Graphics Interchange Format (GIF)**

The GIF format was developed in 1987 by the CompuServe Inc. for the Internet. The GIF format supports colour depths from 1-bit to 8-bits (256 colours) and always stores images in compressed form using lossless LZW compression. GIF is a proprietary format and includes non lossy data compression, making it suitable for on-screen or web use.

#### **2.2.2.1.1.2 JPEG File Formats**

The JPEG format was developed in 1990 by the Joint Photographic Experts Group for data exchange. The JPEG format supports a 24-bits colour depth and uses lossy compression. It can reduce files sizes to about 5% of their normal size, indeed the JPEG is an adequate format to store complex colour images, such as photographs.

#### **2.2.2.1.1.3 Portable Network Graphics (PNG)**

The PNG format was developed in 1996 by the PNG Development Group to provide an open alternative to GIF and to the licensing problems related to LZW compression. It supports colour depths from 1-bit to 48-bits and always stores images in compressed form using the lossless LZ77-based Deflate compression algorithm, an algorithm which isn't patented and is consequently free for use.

#### **2.2.2.1.1.4 Tagged Image File Format (TIFF)**

The TIFF format was developed in 1986 by the Aldus Corporation Inc. for scanning activities and desk-top publishing. The TIFF format supports colour depths from 1-bit to 24-bits, a wide range of compression types and uncompressed data.

### **2.2.2.1.2 Vector File Formats**

#### **2.2.2.1.2.1 EPS (Encapsulated PostScript)**

The EPS format is used by PostScript language. This kind of file usually contains a bit-mapped representation of the graphics for on-screen display purposes and the vector image information for printing, EPS is preferable for print, since it is device independent and provides the best output at any size and resolution.

### **2.2.2.1.3 Hybrid File Formats**

#### **2.2.2.1.3.1 WMF (Windows Metafile format)**

WMF was developed by Microsoft and is very widely implemented on the Windows platform, but is strictly platform dependent. A WMF file consists of a set of Windows specific instructions to draw a vector graphic. This is an excellent format for image interchange between Windows applications, but is not very interoperable.

### **2.2.2.2 Audio formats**

Audio materials are at risk, as their carriers deteriorate and format specific hardware and software become obsolescent. Old and obsolete audio formats are represented by cylinders and coarse groove discs that can be managed only by professionals who have the suitable equipment. Other formats that are close to obsolescence are vinyls, quarter inch tape, micro cassettes, and mini-discs. In the case of audio formats, any preservation strategy must be based on digitalisation. The guidelines to follow when creating and preserving Digital Audio Objects are:

- *Optimal signal retrieval from analogue carriers*: the equipment has to be suitable to replay parameters of the original recording.
- *Unmodified transfer to new target format*: signal must be preserved free of alterations, “improvements”, de-noising etc.
- *Improvements of transfer technology*: when mechanisms for the improvement of the transfer technology are applied, it is anyway necessary to keep the original for possible later consultation.
- *Digital formats*: the recommended digital format is the de-facto WAVE standard.
- *Digital archival principles*: digital preservation copies must be free of irremediable errors or with lowest possible rate of remediable errors. The migration has to be performed before the systems/formats become obsolete and it's better to produce and preserve a reasonable number of identical copies.

In the following sections are described the most popular audio formats.

#### **2.2.2.2.1 WAV**

The Wave file format is Windows' native file format for storing digital audio data. It is one of the most popular and supported digital audio file formats on the PC due mainly to the popularity of Windows. Almost every modern program that can manipulate digital audio supports this file format.

Wave files use the standard RIFF (Resource Interchange File Format) structure which is the storage structure used for multimedia data on the Windows platform. It organizes data in “chunks” which each contains a small header, that describes the chunk type and size, and data bytes. This organization method allows programs that do not use or recognize particular types of chunks to process only the known chunks and skip the unknown ones. Data chunks may contain smaller "sub-chunks" of data.

There are a few types of chunks defined for Wave files. Many Wave files contain only the Format Chunk and the Data Chunk. The former contains information about how the waveform data is stored and specific information such as the type of compression used, number of channels, sample rate, bits per sample and other attributes. The latter contains the digital audio sample data which can be decoded using the format and compression method specified in the Format Chunk.

Note that the Wave format is being used as a 'master' audio format for preservation purposes by some archives and archive projects. The main reason is that it supports the PCM (Pulse Code Modulation) format which is what is of importance to the archives[47].

#### **2.2.2.2.2 MPEG audio Layer-3 (MP3)**

An MP3 audio file is organized in smaller parts called frames. Each frame is independent from each other and it has its own header and audio information. The MP3 format is a compression system for music. The MP3 format helps to reduce the number of bytes in a song without hurting the quality of the song's sound since a technique of compression called perceptual noise shaping is used. It is partly "perceptual" because the MP3 format uses characteristics of the human ear to design the compression algorithm. Indeed, it is necessary to consider that there are sounds that the human ear cannot hear, sounds that the human ear hears better than others and

if there are two sounds, the human ear hears only the louder one. Considering these facts, certain parts of audio streams can be eliminated without significantly hurting the quality of the song for the listener. The relevant parts of the audio stream are compressed using other techniques that assure no loss in compression.

#### **2.2.2.2.3 Audio Interchange File Format (AIFF)**

The "Audio Interchange File Format" was developed by Apple for storage of sound in the data fork of Macintosh files. It has been adopted as a standard audio format by the OMFI (Open Media Format Interchange) group. An AIFF file is composed of a number of chunks. In particular, there are the Common chunk that contains the fundamental parameters of the sound (sample rate, number of channels, etc) and a Sound Data chunk that contains sampled audio data.

Note that the AIFF format is used in several projects for capture and preservation of sound files as AIFF includes structural metadata in the form of channel definitions and time marks [48].

#### **2.2.2.2.4 Broadcast Wave format (BWF)**

The Broadcast Wave Format has been developed by EBU Project Group P/DAPA [54]. It has been designed with the goal to provide a means of exchanging programme material between audio workstations in the form of special files. The metadata included in the BWF has been restricted to the minimum needed to exchange audio data. However it is possible to add more information in order to support the production, post production, broadcasting and archival stages of the life of a audio programme. The BWF uses the basic WAV structure already described in Section 2.2.2.2.1. It only needs two new chunks that have to be added respectively to a linear Pulse code modulation and to a MPEG WAV file to make a BWF file. Specifically, a BWF file includes a special broadcast extension chunk that contains information such as the description of the file, the originator, the reference number issued by the originator, the date and time in which the programme is created and a log of the signal coding.

Note that this format is used in the preservation of audio materials by the Phonogrammarchiv, Austrian Academy of Sciences [55].

### 2.2.2.3 Video formats

There are about fifty analogue formats and fifteen digital formats in the field of video representation. Archival procedures in this field are based on the preservation of the original media and/or on transferring contents. The former option has various drawbacks, the first of which is the well-known problem of the ageing process, which doesn't guarantee the integrity of original media over time. The transfer of content from an analogue format to a digital format is instead recommended, since this type of conversion guarantees a minimal loss or distortion of information and regenerates the original file formats, which is a great advantage.

In the following table, the most common video formats are identified and listed by the team involved in the National Initiative for a Networked Cultural Heritage (NINCH) [52].

<b>Extension</b>	<b>Meaning</b>	<b>Description</b>	<b>Strengths/weaknesses</b>
<b>.mpg</b>	Moving Picture Experts Group	Standards created by the group working for ISO/IEC. MPEG-1: for Video CD and MP3 are based on this early standard. MPEG-2: DVD based on this. MPEG-4: Standard for multimedia on the web. MPEG-7: Currently under development; for 'Multimedia Content Description Interface'.	Good quality and low file sizes. MPEG-1 can take a while to load.
<b>.qt, .mov</b>	QuickTime	Created initially for Macs, can now be used on PCs too. This format is compatible with QuickTime player. Quick Time 4 has streaming capabilities.	Excellent quality, easy capture, widely used, can be large. In Windows the QuickTime player takes up lots of space.
<b>.avi</b>	Audio/Video Interleave	This format is played by QuickView, Windows' Media Player. Replaced largely by MPEG and Windows media.	Large files, very good quality, must be encoded/decoded properly.
<b>.rma</b>	RealMedia	It is a streaming format. Proprietary format that is an equivalent to Windows Media.	Requires RealMedia plug-in.
<b>.wma</b>	Windows Media Video	It is a streaming format. Version 8 offers near DVD performance.	

The most common format in digital archiving is the MPEG, but there also exists another format, the MXF (Material Exchange Format), which was studied specifically for the preservation of video material and is described in the next section.

#### **2.2.2.3.1 MXF (Material Exchange Format)**

The MXF format is a “File wrapper and packaging format capable of encapsulating video, audio, single pictures, etc., plus associated metadata”. It is a standardized format, developed by the Pro-MPEG Forum and the G-FORS Group (EU), and supported by the AAF-Association and by major AV-vendors and institutions. Basically, it is a wrapper for multimedia containers: in terms of video elements, the body of the document can be a MPEG or DV file, or even an uncompressed file. The content of the file is described through three variables: key, length and value. The key defines the data type (i.e. video file, audio file, system data, etc.), the length specifies the size of the contained block and the value is the actual content.

#### **2.2.2.3.2 MPEG formats**

The MPEG video format family is composed of two different formats, the MPEG-1 and the MPEG-2.

The MPEG-1 system combines a plurality of coded audio and video streams into a single data stream. The specification provides a fully synchronised audio and video and facilitates the storage in and the possible further transmission of the combined information through a variety of digital media. This system's coding includes information in the bit stream to provide the system-level functions of synchronization of decoded audio and video. The coding layer specifies a multiplex data format that allows multiplexing of multiple simultaneous audio and video streams as well as privately defined data streams.

The basic principle of MPEG System coding is the use of time stamps which specify the decoding and display time of audio and video and the time of reception of the multiplexed coded data at the decoder.

The MPEG-2 concept is similar to MPEG-1, but includes extensions to cover a wider range of applications. The primary application targeted during the MPEG-2 definition process was the all-digital transmission of broadcast TV quality video at coded bitrates between 4 and 9 Mbit/sec.

However, the MPEG-2 syntax has been found to be efficient for other applications such as those at higher bit rates and sample rates (e.g. HDTV). The most significant enhancement over MPEG-1 is the addition of syntax for efficient coding of interlaced video (e.g. 16x8 block size motion compensation, Dual Prime, et al).

#### **2.2.2.4. Plain text File Formats**

The common denominator of text file formats is the American Standard Code for Information Interchange (ASCII) [45]. There are two forms of ASCII: *standard* and *extended*. Standard ASCII only contains codes for 128 characters. It is highly portable, indeed it is transportable across all networks and capable of being accessed and manipulated on all computers. Extended ASCII is a non-standard format containing codes for 256 characters. The character set includes the standard ASCII character codes and other 128 character codes that are machine dependent. It means that each hardware vendor defines the codes for their own platform and for their

specific purposes such as the specification of special graphic characters or of application specific file format codes. Normally TXT is the 3-character PC extension used for plain text files. Text organized in this kind of files has no formatting and is very portable. Indeed, the text can be displayed by any applications.

### **2.3. Emerging Standards**

In this section, new file formats specifically created for preservation needs are studied.

#### **2.3.1 PDF-Archive Initiative**

It specifies a subset of PDF tags for archival purposes as an ISO standard [20]. PDF is a digital format that electronically reproduces the visual appearance of documents, whether they were originally created in PDF, converted from another electronic format, or digitized from paper or microfilm. PDF is widely used within companies, government agencies, libraries, archives, and by other institutions and individuals all over the world. This format is mainly used to collect and disseminate information over the Internet, and store electronic records. As a result, large portions of relevant information are maintained in PDF.

PDF cannot be used as an archival format. Long-term solutions are needed to keep digital PDF records accessible for a long time length. The PDF/A format was expressly introduced for the purpose: it identifies the subset of mandatory, prohibited and recommended components and specifies how the software should deal with these components to obtain the needed file. The components in which technical requirements are expressed are: File Structure, Graphics, Fonts, Transparency, Annotations, Actions, Metadata, Logical Structure, Forms.

In order to satisfy preservation criteria the PDF/A attempts to achieve the objectives of device independence, self-containment, and self-documentation. Self-containment is defined as the degree to which a PDF/A file may contain all the necessary resources for performing interpretation and rendering in a reliable way and as expected to, while self-documentation is defined as the degree to which a PDF/A file would document itself in terms of descriptive, administrative, structural, and technical metadata.

### 2.3.2 XML Working Group Initiatives

It specifies a subset of XML tags for archival purposes. According to D.R. Miller [38], XML (eXtensible Markup Language) is now accepted as the universal format for data and document exchange, and has actually become the *lingua franca* of the Information Age. Currently, 'library information' is suffering from the pace of evolution in the World Wide Web. Universal standards adopted by XML enable the handling of diacritics, special characters, and of non-Roman data like ordinary text, both within documents and computer operating systems and applications. This is something crucial for libraries, and might be strategic in the internationalisation of data networks. Moreover, XML shows the great promise of data longevity (or future-proofing), in a situation in which hardware, software, and network protocols continue to change.

In [39], authors state that XML and PDF are often presented as the two rival formats, with the idea that if you wish to preserve a record on the long-term you should choose one of the two. The other two standards named in the Regulation, TIFF and SGML, end up paying the price of this ideological confrontation. But the truth is that PDF and XML are complementary to the point that, in terms of preservation, it is actually better to use them both, rather than choose one of the two. And actually, choosing both standards is also a way of 'sharing the risk': if within a hundred years one of the two formats will no longer be readable, the other might still be so. The ideal solution would be that of developing another open standard to replace the PDF, so that the safekeeping of our digital inheritance would not be left in the hands of a single company.

---

### 3 Classification of file formats properties

In this section different classifications of file formats widespread in literature will be examined with reference to a sort of assessment analysis based on various characteristics and aims. A first classification defined in [26] in which file formats are simply distinguished between:

- **proprietary format:** the proprietary formats are licensed and the full documentation is not always available. The user cannot modify the format freely.
- **open format:** open formats are always fully documented, they are not licensed, and the user can freely modify the format structure.

In the next sections classifications based on multiple criteria are presented and discussed.

#### 3.1 Characteristics that can be used to classify and assess file formats

The characteristics that can be used to classify file formats are sometimes mentioned shortly in the literature. An important contribution is provided by a document released by the European Commission (IDA group) that introduces a classification that defines the characteristics of file formats [1]. It differentiates characteristics as technical and not technical aspects. The technical aspects considered are the following:

- *Openness:* the requirement to define a format as an “open” standard is that the document format is completely described in accessible documents. The description has to be published and distributed freely and the programs that use the document format do not have restrictions, or legal bindings. The file is open for self-made modifications.
- *Non binary:* The document can be saved as binary data stream (i.e. Word, PDF) or in plain text format (i.e. XML and RTF). There are some limits on binary format such as *platform dependency* and *increasing awareness for long-term archiving*. The former implies limitation on the use of heterogeneous architectures and applications. The latter is critical for preservation, indeed with the binary formats it is not guaranteed that the

conversion of old formats in newer ones let access data at all times. Currently, considering the rapid change on document formats, there is not satisfactory solution available based in proprietary document format for documents that are required to be archived.

- *Modifiable*: the document can be modifiable or not modifiable. The modifiable formats are used for collaborative projects in which one or more recipients can modify the information. In contrast to these formats there are the not-modifiable formats in which it is possible to share information but not modify it. Documents of this last type are used only to transmit the contents of a document keeping the format. In the preservation processes, the not modifiable contents should be preferred since it is important to preserve the representation of the document over time without changing it.
- *Format Fidelity*: The preservation process is efficient if the documents do not lose meaning and value and if the layout or visual emphasis is altered. Formally, it is possible to distinguish among presentation, content and structural fidelity. The presentation fidelity of a document format is defined as its ability to preserve the original layout of the document, regardless on the platform used to open it. The presentation fidelity can be classified as:
  - *High*: the document layout is preserved across platforms and computers
  - *Medium*: there are important issues associated with the document format. This can be due to architecture dependency or incomplete documentation of the format (i.e. word)
  - *Low*: the document layout depends on the user's viewing preferences (i.e. HTML)

The presentation fidelity is often mainly dependent on the used software platform and can be argued that this is a software matter and not really related to the file format. The content and structural fidelity are more strictly to the file format and the fidelity can be expressed through the concepts of:

- *Accuracy*: The accuracy is the extent with which the digital document content and structure are correct along the initial specifications.
- *Reliability*: The reliability is defined as the degree with which a document can be trusted to convey the right information.

The format fidelity maintenance in all its aspects is an absolute condition that cannot miss in a preservation activity.

- *Cross-platform interoperability*: the interoperability implies that the format can be accessed on various hardware and software platforms in its representation and its semantics.

Non technical aspects here considered are:

- *Functions*: definition of the file format along with what it is possible to do with the format itself.
- *Roles*: effective use and sustainability of format.

Another important contribution is provided in DELOS NoE Deliverable “A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies” in which Andreas Rauber and Carl Rauch propose a new hierarchy. They propose to differentiate the criteria in two different levels. The former includes the *file characteristics* such as *Appearance*, *Structure* and *Behaviour*. The latter includes *process characteristics* such as *Authenticity*, *Stability*, *Scalability*, and *Usability*. Note that the process characteristics are more difficult to describe than the file characteristics. Finally, it is also considered a third level in which costs are listed. In particular, *technical* and *personnel costs* are considered.

For each identified characteristic, sub goals are defined. Sub goals in some cases are strictly related to the file type. For example, as concern the appearance criteria that refers to the visual impression, sub goals as “page” and “paragraph” have been identified for documents, while bit rate and frame rate are considered the sub goals respectively of sound and video category.

### **3.2 Criteria for selection of suitable file formats for digital libraries**

After a general discussion on file formats and their classifications, it should be relevant to identify some criteria – as discussed in the related community - for the selection of suitable file formats for digital libraries. Each criterion will be here presented together with the specific requirement that satisfies in digital libraries.

In the digital library literature there are many contributions that suggest which are the file formats more appropriate for preservation issue. The preferred formats should be those that remain usable for a significant amount of time. The set of preferred formats

should be strictly related to ensure the preservation of the resources, but should also be determined in order to include the various file categories necessary to archive all the different aspects evaluated as relevant in this area. In [15] the author states that the main dimension to consider is the openness of the format that is its independency from providers or computer applications. Four types of basic file formats are considered within the digital library communities: text, image, sound and video and for each categories specific standard format are suggested as described in the table below:

<b>Type of file</b>	<b>Format suggested</b>
Text	Unicode (ASCII), XML and PDF/A
Image	raster: standard TIFF for master copies (no-compression, high resolution), JPEG for safety copies or distribution vector: CGM, EPS, DXF, SVG
Sound	compressionless WAV (PCM-coding)
Video	MPEG

The choice of a file format is necessary to consider the possible future re-use. Clausen L.A [29] proposes a methodology to evaluate the reusability of a file format and the degree of obsolescence. The suggested criteria – here presented - are defined on the basis of their capability to be used for archival purposes:

- *Openness*: formats that are described by publicly available specifications or open-source code can, without a significant effort, be rebuilt at a later time. As regards specifications, open and publicly available specifications allow to build viewer even if the original one is not available anymore. The freeness of the specifications from patents or copyright make easier to have free viewers for a medium-long term preservation. Openness can be also considered as regards the source. If the source is freely available, this requirement allows its replication in case the original cannot be compiled. It is sufficient for a viewer to be under the GPL license [35] to ensure its continuing freely availability. Finally the format should not be encrypted because the use of a special encryption key to read the file is at risk of obsolescence, as the key may be lost as well.
- *Portability criteria*: a format that is defined in an independent way will be much easier to use in the future that a format that works with a specific hardware or software platform. Indeed, in particular hardware dependency is

dangerous in a format because hardware changes fast. Operating system dependency is also restrictive even if operating systems change slower than hardware. As regards extra-software requirements, the format can be considered stable along with the software involved in its use. In this area it is important to consider the degree of usage of a format. A widespread use of a software indicates that others consider it useful and important. It also means that there is an interest to create viewers and tools in the future. Finally for portability criteria the format should not be available in many versions since more versions of a format is available, more difficult is to understand it, in particular if the differences between versions are not really clear.

- *Quality*: the quality of a format indicates the degree required for accomplishing a task at the present. Quality is a multi-dimensional concept that takes into account many aspects often in trade-off among them. Clausen L.R. [29] considers:

- “low space cost”: smaller files are easy for tools to handle;
- highly compassing: a format that can be used as a target for a greater number of other formats saves resources otherwise necessary to maintain other formats;
- Robustness: a property that guarantees from random bit errors and loss of parts of the file; compressed formats are vulnerable to bit errors;
- simplicity: if a format is simple, more likely new viewers can be created in the future that will handle the format correctly;
- highly tested, widespread and long-term use gives us more assurance than a format is of high quality;
- loss-free: converting a loss-free format into a lossy format implies that some information will obviously be lost;
- supports metadata: it may allow us to gain metadata about the source of files that would not otherwise be available, and also provide a redundancy of metadata in case the externally metadata are lost.

Christensen S.S. published a document [37] focused on archival data format requirement defining in details all the features that a file format has to satisfy to be considered suitable for archive purposes. First of all, the basic concept of openness,

portability and quality analyzed above are proposed. In addition the following requirements are required:

- A basic specialized requirement for the openness feature is that the *file format has to be OAIS compatible*. File format has to match the requirements of the OAIS model [41].
- *The format must support all important Internet protocols*. A protocol is supported if it is possible to store and retrieve all relevant information transferred by the protocol.
- *The format must support metadata*. The author considers crucial to have the possibility to enrich the stored bit-stream with metadata.
- *Data integrity has to be guaranteed*. It has to be possible to detect bit errors inside archived data and to easily correct them.
- *Data backup must be simple*. A simple and reliable backup procedure must exist for the selected format.
- *The format must support authenticity information*. It must be possible to identify the person responsible for adding a collection of data to the archive.

*Transformability* is another important characteristic of file formats that has not been considered in earlier classifications discussed above.<sup>2</sup> It relates to their self checking capability, which makes them “*transformation safe*” to a higher or lesser degree. When a file cannot be damaged without an automatic process accessing it becoming aware of the damage (even if it can not be recovered), it can be trusted, that all successful automatic conversions of that file will result in new valid files. If that can not be guaranteed, any processing requires considerable effort to check, whether the file was actually still valid when being transformed. Some metrics to evaluate a file format along these lines suggest themselves immediately - e.g. the maximal percentage of information that can be corrupted without such corruption being detectable. To build an implementable set of rules from such metrics, considerably more research into the processing capabilities provided by formats will be needed.

---

<sup>2</sup> The concept of Transformability arose as part of the work by Cluster (WP6) at its meeting on the 21<sup>st</sup> and 22<sup>nd</sup> of October 2004 in Cologne.

In the next section all the criteria presented will be used to analyze the attitude of specific file formats to preservation

### 3.3 File formats frequent use and preservation

Considering the document-like and graphics formats introduced in Section 2.2 and the characteristics for preservation presented in Section 2.3 in this section we intend to discuss the attitude for preservation of the most spread file formats.

#### 3.3.1 Document-like formats

Document-like formats are used to manage text and eventually other graphics formats. The readability of a digital document is guaranteed if all the text contained in it is readable. Readability can be ensured by both *Word Processing tools* and *Viewing tools*, since they are able to manage the presentation layer. More problems are related to the comprehensibility aspect. Indeed, meaning could be added to the text by using mark up features. Mark-up mechanisms are not supported by all file formats in this category. Mark-up features are needed to preserve the presentation layer and all the graphics elements contained in the document with the text. Further, the file format has to be able to add metadata and handle them. Finally viewing tools are not able to preserve functionality, since they are not focused on to the management of the internal structure of the document.

Features	XML	Microsoft DOC	RTF	POSTSCRIPT	HTML	ADOBE PDF
Openness	+	-	-	+	+	+
Portability	+	-	-	+	+	+
Quality - Highly compassing,	+	-	-	+	+	+
Quality - Robustness	+	+	+	+	+	+
Quality - Simplicity	-	-	-	-	+	-
Quality - Highly tested	+	+	+	+	+	+
Quality - Support metadata	+	+	+	+	+	+

It is possible to state that XML is the best format as regards the Word Processing tools for the preservation. In the digital world XML has acquired a solid position it is of the

greatest importance for digital preservation, not just because of this widespread use, but mainly because it is highly portable and open. Indeed, it is both platform- and software-independent. The separation of content, structure and appearance plays an important role here. Because much of the software-dependency is associated with the appearance or form (for example on the software of Adobe in the case of PDF), the chance is much greater that the abstracted content in XML can withstand the course of time. PDF is a suitable format managed and produced by Viewing Processing tool. Often XML and PDF are put forward as two rivals from which one must choose in order to preserve a document for the long-term. PDF and XML are complementary, and it is actually more appropriate to use both XML and PDF for preservation of a document than to choose between XML and PDF [39].

In [53], the most important preservation strategies are compared and authors found out that the choice of the right strategy mainly depends on the file format. The Utility Analysis has been used to compare the alternative strategies. An example shows the application of the proposed metrics on the preservation of Word 2002 files. The considered alternatives are the migration to MS Word 2003, the migration to the OpenOffice Writer format, the migration to PDF and the choice to not make any changes. Alternatives have been compared according to the criteria of the objective tree presented in Section 2 of the Deliverable "A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies" [53]. Results show the ranking of alternatives that consider the migration to MS Word 2003 as the most suitable solution. A good solution is also the migration to PDF. The migration to Open Office and the choice to not make any changes are not acceptable. In the specific example, it is interesting to highlight that PDF gets lower scores than Ms Word for some criteria such as structure and word functionality. For details on the different criteria please refer to [53].

### **3.3.2 Graphics File Format**

The long term preservation of image is based on procedures, tools, standards, specifications and guidelines available to realize the long term access of digital images. The three main building block of long term preservation are:

- *Standard graphics file formats*: the main characteristic of standards is the durability
- *XML data format*: Digital data encoded in the XML data format is durable data
- *Metadata* on digital objects is essential in order to understand and process digital objects in the future,

Standard graphics file formats are characterized by the following characteristics:

- They are used by large community during a considerable period of time
- Specifications must be available in the public domain or published by SDO
- Wide range of systems has to support the format
- Support of uncompressed files is required
- They must contain facilities to store preservation metadata
- They must enable coding of all significant characteristics of analogue original

The most spread standard graphics formats are TIFF, GIF, JPEG and PNG format. The comparison among these formats along the characteristics listed before has shown in Table 1.

<b>Features</b>	<b>TIFF</b>	<b>JPEG</b>	<b>GIF</b>	<b>PNG</b>
Used by large community during a considerable period of time	+	+	+	-
Specifications must be in the public domain or published by SDO	+	+	+	+
Wide range of systems has to support the format	+	+	+	+
Support of uncompressed files	+	-	-	-
Facilities to store preservation metadata	+	-	-	+
Enables full information capture	+	-	-	+

As shown in the table the TIFF format is considered the most suitable format for preservation. Indeed, in many initiatives the TIFF format is used as standard for preservation and is the final format for migration patterns as regards graphic files

Other initiatives focus on the use of XML as language to describe digital images. There are specific expressions to identify the content model in XML in terms of elements and attributes that are part of the bit stream, e.g. standardized colour coding of pixels. The definition of the image in XML enables the preservation supporting the migration from an existing format to an innovative one.

The third building block for preservation of image file formats is the use of metadata. Metadata can be stored in digital images in three ways:

- as part of the image (e.g. File header)
- in separate database
- in file system

The preservation metadata can be created from scratch or (re) using existing data elements

### **3.4. Analysis of the main projects in this area.**

As discussed in [23] five spread approaches used by information systems to face the proliferation of file formats and available information have been identified:

- *Specialized niches*: applications define a proper rich format and manage with different applications the conversion with the most common formats. This is the case of proprietary formats that scarcely support heterogeneity and interoperability.
- *Lowest common denominator*: applications that want to share data widely may use simple formats such as ASCII, or HTML. The disadvantage of this approach is the poor expressiveness of the simple format in which there is not the possibility of using rich data type.
- *Simple data type schemas (Format Registries)*: data type schemas involve central registries that define set of formats. The type registries used today are not expressive enough.
- *New standards*: new formats can be sometimes so widespread used to be defined and publicized as standards.
- *Polyglot client applications*: the information system in this case is provided with interfaces and tool able to understand and manage a huge variety of file formats.

In this section a specific emphasis will be done on the critical aspects of these projects as related to the digital preservation.

### **3.4.1 Simple data type schemas(Format Registries)**

#### **3.4.1.1 National Archive Initiative**

The National Archive Initiative addresses the secure storage and operational management of records in a service environment. PRONOM, the file format reference database is the key component of this digital archiving work [19]. PRONOM is an on-line information system about data file formats and their supporting software products. Originally developed to support the accession and long-term preservation of electronic records held by the National Archives, PRONOM is now being made available as a resource for anyone requiring access to this type of information. PRONOM holds information about software products, and the file formats which each product can read and write. PRONOM is a file format registry. This is build in order to apply different classification schema to file formats. The most important propriety of the file PRONOM format registry: is the persistence to act as a resource which a digital repository can actually point to. The unique identifiers allow the unique retrieval of the digital repository. A file format registry needs to be wide to include all the major file formats in the usage. The unique identifiers have to be unique, unambiguous and persistent

#### **3.4.1.2 Global Format Registry**

An initiative Harvard Library Digital Initiative (LDI) [17] and MIT DSpace [18] projects aims to the realization of a centralized global format registry for the wider digital library community [16]. The data model for the registry includes four categories of information to associate with the file format [16]:

1. *General descriptive properties*, including canonical and alias identifiers for formats
2. *Characterization properties*, detailing the syntactic and semantic properties for formats
3. *Processing properties*, describing systems and services for which registered formats are inputs or outputs
4. *Administrative properties*, capturing important events in a registration's provenance

### **3.4.2 Polyglot client applications**

#### **3.4.2.1 TOM (Typed Object Model)**

TOM [44] is a data model for describing a wide variety of data type and formats in well-defined, machine-processable manners. TOM is composed of a distributed system of “type brokers” that maintain and interpret these descriptions, and operate on data in the formats they describe. It works with existing data formats, systems and there are not prerequisites for formats other than being byte sequences. It supports two preservation strategies such as format migration (with controlled information loss), and functional emulation.

In TOM, a file format can be defined as a type combined with a sequence of encodings that represents the type. A type describes the information contained in an object (a unit of data) in terms of:

- What is contained in the object (attributes): e.g. a URL has a protocol part, a host/port part...
- How the object “behaves” when interpreted (operations): e.g. a URL can be resolved to yield another object
- What constraints exist on the object (semantics): e.g. a URL’s port number must be non-negative

#### **3.4.2.2.FCLA (Florida Center for Library Automation) initiative**

FCLA (Florida Center for Library Automation) initiative [40] provides a series of specifications for preserving specific file formats. In the web site some of the preservation actions are described. “The FCLA Digital Archive will accept digital documents along with the appropriate metadata, and safely store on-site and off-site copies of the files. An action plan will be developed for each file format, which might include the creation of canonical derivative versions and/or format migration. A primary characteristic of the FCLA digital archive is its role in serving the real needs of a diverse group of libraries. Every effort will be made to accommodate the formats important to the institutions, whether these are traditionally considered "archival" or not.” The main task in the adopted preservation strategy is the decision of how 'stable' a format is and what the distinguishing characteristics of a format are. An action plan is defined for each file format in order to describe the long- and short-term actions that will be taken.

## **4 Relationship between file formats and digital preservation**

### **4.1 Introduction**

The main goal of the preservation activity is ensuring long-term access to digitally stored information[2]. In this process file formats play a critical role because the long-term access can be assured only if the used file format can be accessed over time. But before considering which is the most appropriate preservation technique, it is necessary to define what aspects of digital documents are important to preserve. In theory, it is not possible to define a unique model to adopt for the document preservation, but it has to be considered the context. Indeed, in different context could be more important to preserve the look like of the objects and in other cases can be fundamental to preserve functionalities. It is absolutely relevant to identify the specific characteristics of each typology of digital resources to be preserved with reference to its function: an archival resource requires different criteria and elements to be considered with reference to the preservation of a single entity like an electronic book. It should be also added that the percentage of these peculiar qualities and nature is changing in favour of a lower degree of differentiation, which in any case is still persistent and requires to be considered specifically when discussing the elements to be preserved. A too general perspective is useful to share tools and awareness, is dangerous if it implies a low level of understanding.

### **4.2 Basic file format requirements for preservation**

In the literature many authors try to define which are the aspects that have to be preserved.

The basic requirements for preservation are:

- *Context*: all the information about the context in which the preservation process is performed has to be saved. Information about the context could be general as the name of organization, and business processes involved, or more detailed as the single relations with the other documents. In this phase, it is also necessary to maintain a preservation log file with information about original and current file formats, name and versions of software, hardware and Operating System and the preservation actions.
- *Content*: all content must be preserved, including headers, footers, table of content, document properties, remarks.

- *Structure*: Structure of the document must be preserved, in order to represent the logical relations between the components of the document, such as the order of chapters, paragraphs, but also the right position of inserted remarks, footnotes and images.
- *Appearance*: The appearance of the original and the preserved version do not have to be identical, but the new appearance may not in any way affect the meaning of the original record.
- *Behaviour*: description of active links must be preserved.

In [29] the author considers five fundamental aspects for the preservation. These aspects are more oriented to the quality of the preservation and are listed in the following:

- *Readability*: The most important function is that the object can be read. For documents with text, it is easier because a simple text extractor can interpret the text. Complexity increases with images, sounds, and movies. Many techniques suggest to preserve significant parts of the complex objects able to describe the whole content or the main part of it. This is the main factor if the preservation context is interested purely in the content of the objects to preserve. Indeed, this aspect is the least costly aspect to preserve but has the greatest amount of information loss.
- *Comprehensibility*: in many documents the effort to understand the text is more complex to follow line of words. The reason is that the text is not organized in lines but it could be contained in tables or there could be symbols to emphasize the meaning of the text. The structure or added elements could be also important and the loss of these information means the loss of the comprehension of the whole text. This aspect is easily preserved for images and sounds where some errors could be introduced but it is hard to interfere with easy comprehension. The same thing is valid for the movies where even if the resolution is low it still viewable as a movie.
- *Appearance*: This is the only aspect in common with the above classification. This element is rarely essential. However, a good rendition of the files adds to the confidence in the data and gives a better impression of what the original document looked like. In some context this aspect is very important, especially if it is

important to have the image of the original document. For example, the appearance is very important in the context of art or history research.

- *Functionality*: in digital objects many functionalities are hidden such as formulas or hyperlinks and the preservation strategy adopted can caused loss of information. There are context in which there is no sense preserving objects without their associated functionalities.
- *“Look and feel”*: there are contexts in which is not important to preserve only the appearance or the functionality but all the look and feel of the object.

The choice of the preservation strategy have to take in account the most important aspects in the considered context. Most of times there are trade offs among the aspects to consider and it is important to assign a certain priority degree to each aspect. This is necessary to establish the most suitable preservation strategy along the context requirement. Further, the big deal is facing the trade off between how much it is possible to preserve and the resources organization can spend to preserve them. Clausen L.R [29] states that the preservation of the more widespread formats must take higher priority, but at the same time, the most widespread formats are the ones most like to have viewers in the future. The most problematic formats may well be ones that are widespread enough that losing them would lose significant amounts of data, but not widespread enough that we can feel sure that somebody will always create be there to create a viewer. Always Clausen provides an interesting comparison of the formats along the considered aspects.(Table below)

Category	Readability	Comprehensibility	Presentation	Functionality	Look & Feel
<b>Document-like</b>	Text	Text with some markup	All markup and graphics	Links work	
<b>Image</b>	Low resolution	Medium resolution	High resolution	-	
<b>Sound</b>	Lowest bit rate/sample space	Medium bit rate/sample space	High quality	-	Includes player
<b>Movie</b>	Some images	Low resolution, some artefacts	High resolution, few artefacts	DVD menus work	Includes viewer
<b>Data</b>	Text extract	Text in columns	-	-	-
<b>Structured graphics</b>	Text extracts	Image capture	Vector format	Connections, checks etc	Same interface
<b>Spreadsheets</b>	Text extract	Text in correct positions	Correct text and graphics	Formulas	Same interface
<b>Collections</b>	Index	Separate files	Archive	-	-
<b>Configuration</b>	Text extract	Structured files	-	-	-
<b>Programs</b>	Screenshots or film	Semi-functional emulation	-	Full emulation	Program runs

Through the observation of the table above it is interesting to note that each category of file format is characterized by different aspects. The choice of the most suitable strategy should be done considering these aspects and their relevance inside the specific context. The context is defined mainly by the type of files to preserve, the nature of the resources and the purpose of the preservation activity.

### **4.3 Preservation strategies**

This section describes the current preservation strategies. It is possible to distinguish preservation strategies between technology preservation strategies and information preservation strategies. The technology-oriented strategies consider that digital resources can be stored on any medium that is able to represent their binary digits or bits. The first requirement for the digital preservation is the stability of the medium [3]. The medium has not to become obsolete before the information is transferred on another medium. In order to predict the obsolescence of a medium, there are studies that identify the lifetime periods associated with media to help the preservation activities [4]. The lifetime is defined as the time period in which the information stored in a medium can be accessed and retrieved without any losses. The technology oriented strategies include technology preservation and technology emulation.

The technology preservation is not sufficient when the information cannot be accessed and retrieved because the file format in which it is encoded is obsolete and not manageable by any program. In this case, it is necessary to transform or convert the old file format in a format that is independent by the hardware or software that created it. Information migration and information encapsulation are the most important techniques in this field.

The choice of the preservation strategy has to be based on specific requirements dependent on the different file collections and on the context in which the preservation is performed. The selection phase can constitute a critical matter for an organization for the difficulty to find comparable elements among the different strategies and to measure the degree of a strategy to satisfy the expressed requirements.

Christensen S.S. [37] suggests another possible categorization of preservation strategies. It is possible to distinguish between bit-preservation and logical

preservation. The purpose of bit-preservation is to ensure that the collected data are preserved. The purpose of logical preservation is to ensure that it is possible to understand the preserved bits. The archival storage format must accommodate the storage requirements of both bit- and logical- preservation. It is important to add additional metadata to the stored bits in order to ensure the long term preservation of a data collection.

### **4.3.1 Technology-oriented preservation strategies**

#### **4.3.1.1 Maintenance of the original environment (technological preservation)**

This strategy focuses on the preservation of the original technologies to access old formats. All the obsolete equipment is maintained in order to recreate the old configurations in terms of both hardware and software [5]. Operating systems, application programs and hardware are preserved. This strategy is a good short term solution for preserving materials but it can not be applied as a long-term or permanent solution because of the high maintenance costs and the unavoidable aging of the materials. This is a consideration that is also stated by Clausen L.R. [29] that suggests not to apply the technological preservation for the volatility of physical machines. They are subject to wear and tear. The most vulnerable parts are the mechanical parts like drives and fans, but chips can also start failing after a short time. An additional problem is the physical space required to store the machines and the consequent selection of the machines to preserve. Further, the serious failure of the hardware preservation is that the objects no longer can be considered entirely digital. The hardware required must be considered a part of the archive and restricts the options available for backup, refreshing and viewing.

#### **4.3.1.2 Emulation**

The emulation regards only the software environment and consists in the reproduction of the behaviour of the old platform on a new technology. This technique saves the look and feel of the applications as well as their functionalities. The mechanism is based on a virtual machine on which the emulator runs. The problem moves to the migration of the virtual machine on different platform [6].

On this topic there are contrasting opinions in the literature. For example, in [7], the author writes that the emulation has the great advantage to preserve the look and feel of the applications. This advantage is compensated with the disadvantages of complexity of the emulators due to the technology evolution. He concludes that it is a

partial digital solution. On the other hand many authors think that emulation offer the best solution for very long term preservation for resources for which the future use of the material is doubtful [8], or for complex digital resources such as executable files [9] [10][3]. Finally, Clausen L.R. [29] states that emulation is a beneficial strategy for those people that are interested in preserving the presentation and functionalities but it is nor recommended for people that are only interested in readability and comprehensibility of the original document. The main disadvantage of this strategy is also the big effort required in the software maintenance.

### **4.3.2 Information-oriented preservation strategies**

#### **4.3.2.1 Migration to analog format**

Clausen L.R. [29] suggests as simplest preservation strategy the transformation of digital objects in non-digital objects such as prints or microfilm. This media will surely survive and there is a large body of research on preservation of such media from several hundred years of archiving in libraries. The two main problems with this approach are indexability, that is the ability to find and retrieve documents timely and the loss of functionality. There are two ways of capture: the analog capture and the digital capture. As regards analog capture, indexability is assured since documents can be indexed at their creation, even if their retrieval would be dependent on manual intervention, leading to access time that can be several hours or even days. Functionality is all the things a digital object can do that an analog cannot such as hyperlinks, hidden formulas or animated behaviour. The digital capture consists of transformation of the digital object in a much simpler one. The big advantage here is the easy searchability but there are the same functionality problems of the analog capture and additionally there are problems about the existing of media that may itself become obsolete.

#### **4.3.2.2 Migration**

Migration strategy regards the migration of records from obsolete formats to newer ones and the transfer of current hardware/software platform to another [11]. Migration has the advantages to preserve integrity of digital data, retain ability to retrieve/access/use data and exploit technological progress. The disadvantages of this technique are that the exact digital copying is not always possible, the compatibility

maintenance is not guaranteed and the migration procedure is time-consuming, costly, complex and error-prone.

The Open Archiving Information System (OAIS) model considers four categories of migration: refreshment, replication, repackaging and transformation [12]. Refreshment ensures that a reliable copy of the bit stream is maintained while replication and repackaging ensure that a usable file is available. Finally, transformation modifies the bit stream of a digital object. This is the most important meaning in the migration field.

[2] discusses on the different typologies of migration paths. The *continuous* migration path assures that the software is backward compatible. The *standard formats* approach consists in migrate files from the great multiplicity of standard to a smaller manageable number of standard formats. The problem is the choice of the final standard that depends on the structure of the digital source, the objectives and on the user requirements.

A report of the Commission on Preservation and Access (CPA) points out that there are a large variety of migration strategies: minimum migration, preservation migration, recreation, human conversion migration and automatic conversion migration.

Many authors consider migration as the most promising preservation strategy for the future. [8] thinks that migration is a suitable strategy for short or medium term.

Recently, XML is the widely accepted standard considered in the migration and it is successfully used in the preservation field. In the following, two of the most spread migration forms are discussed in details: the migration to standard formats and the migration on request.

#### **4.3.2.2.1 Migration to standard formats**

Standards define the way to use a specific file format for long-term preservation of electronic documents. This approach has the advantages that action can be taken early to preserve data, and it requires no long-term maintenance of special programs nor significant changes to the presentation system. The main disadvantage is the dependence on the method used to convert the file format that may be obstructed by the complexity of the file or lack of information about the file format. Further, the conversion is strictly dependent on the context and especially on what the user wants to preserve. In some cases it is sufficient to preserve the text and in some cases it is

required the most exact conversion possible. The spread approach in the conversion field is performing several conversions of a given file [29]. Indeed, it is possible to convert the file in a simple format to preserve the text and a more complex conversion should be applied to preserve the look and feel of the document and it could be performed in a less reliable format. In this way, it is possible to avoid the all-or-nothing scenario of preservation. The conversion in simple format enables the use of a large number of tools for search and analysis. Simple files can be managed in a simple and efficient way. Note that the conversion method implies some loss of information. Different formats contain different selection of possible information, for example it is sufficient to think about trivial matters such as colours or shapes, they are managed in different ways in the different formats available. Further sequential conversion of same file accumulates errors. It is always better to perform conversions considering the original file.

#### **4.3.2.2.2 Migration on request**

The Camileon Project [24] (Section 3.4.2) suggests an alternative solution to sequential migration, namely **migration on request** [36]. This strategy is based on a set of conversion utilities that are maintained indefinitely, and conversion from the original data into the best format available is done at access time. It avoids the problem of accumulated errors but implies more management costs due to the conversion utilities maintenance. Indeed, the conversions tools have to be kept and modifiable in the future. Commercial tools may do the job at present but it is not guaranteed that the company will continue to produce and maintain them. A possibility is to build and maintain conversion tools but it is a very difficult task.

#### **4.3.2.3 Modification of the file format for preservation**

This preservation strategy consists of acceptable modifications that guarantee the integrity of the documents and that enable its preservation (i.e. adding metadata etc.).

#### **4.3.2.4 Encapsulation**

Encapsulation allows combining several elements to create one unique entity. File formats are maintained in their original form. This technique allows identifying documents as opposed to using them.

Encapsulation can be achieved by using physical or logical structures called containers or wrappers to provide a relationships between all information components such as the digital object and some supporting information including metadata.

Two well known example of encapsulation are the Universal Preservation Format [13] and Rosetta[14]. The former is based on a wrapper that holds the digital object and the metadata together to protect against technological obsolescence. The latter is a method for storing the representation information needed to interpret the digital content of an object separate from the encapsulation to avoid duplication effort and inefficient use of the storage space.

#### 4.3.2.5 Filming

Clausen L.R. [29] considers filming a valid preservation strategy for complex objects since it allows the preservation of the usage and look and feel by filming a user using the file. The filming could be done using the technique of continuous screen capture or the physical filming, possibly including comments. Filming is considered one of the most resource-intensive forms of archiving. It requires a huge number of resources, both physical and human. Further, another disadvantage is the size of the final file that could be significantly larger than the original to preserve.

#### 4.3.2.6 Summary of risks and benefits

In the following table there is a comparison about benefits and risks of the considered preservation strategy:

Strategy	Risks	Benefits	Resource usage
<b>Technological preservation</b>	Mechanical failures, very impractical to use	Perfect look and feel	Very high: high maintenance costs and the unavoidable aging of the materials
<b>Technological Emulation</b>	All-or-nothing preservation, may be impractical to use, software dependency	Near-perfect look & feel and functionalities, no conversion errors.	High: create and maintain emulators
<b>Capturing with “Flat” formats</b>	Loss of functionality, impractical to use	Format independence	High: Manual handling and indexing
<b>Migration</b>	Conversions errors, the compatibility maintenance is not guaranteed	Preservation of integrity of digital data, retaining ability to retrieve/access/use data and Exploiting technological progress	High: Migration procedure is time-consuming, costly, and complex
<b>Migration to standard formats</b>	Cumulative conversion errors, some loss of functionality	Independent of software, Multi-aspect conversion possible, viewing is easy	Medium: Monitor format evolution, obtain and use converters when needed
<b>Migration on request</b>	Conversion errors, some loss of	Viewing is easy, fewer errors than early	High: create and maintain converter suite

	functionality, software dependency, viewing delay	conversion	
<b>Modification of the file format for preservation</b>	Addition of complexity in metadata management	Preservation of the original file format	Medium: resources needed to update and monitor procedures to manage metadata
<b>Encapsulation</b>	Loss of functionality, software dependency	File formats are maintained in their original form. This technique allows identifying documents as opposed to using them.	High: maintenance of the physical or logical structures called containers or wrapper
<b>Filming</b>	Loss of functionality, impractical to use	Shows functionality and actual use, format independence	Very high: Run recording sessions

## 5. Conclusions and future work

The file formats analyses pointed out that XML and PDF are the preferred languages for the migration and conversion strategies. These two languages are often considered as two rivals from which one must choose in order to preserve a document for the long-term. PDF and XML are complementary, and it is actually more appropriate to use both XML and PDF for preservation of a document than to choose between XML and PDF. Future work could consider the analyses of the combined use of these languages and its risks and benefits. About XML and PDF, another question that needs an answer is the real suitability of these languages to preservation. It is possible to be sure that they have the characteristics to last for centuries?

Furthermore, the study the specific context in which the preservation strategy is adopted could help to identify which aspects of the document are important to be preserved and along them determine the suitable preservation strategy. Future work on this field will provide a methodology to identify the suitable preservation strategy along the context and the aspects that are considered relevant in it.

Finally, another aspect that is not deeply studied in this field is the metadata management for preservation with the consideration of technological and costs constraints.

## 6 References

- [1] Valoris, "Comparative assessment of Open Documents Formats Market Overview" – Specific Agreement n. 3 – IDA. 20030523. Available on line at <http://europa.eu.int/ISPO/ida/export/files/en/1928.pdf> (Last access: 29th July 2004)
- [2] Lee, K.H., Slattery O., Lu R., Tang, X., Crary, V. "The State of the Art and Practice in Digital Preservation".
- [3] Rothenberg, J. "Ensuring the longevity of Digital Documents", *Sci. Amer.* Vol. 272, No. 1, pp. 42-47, 1995.
- [4] Zwaneveld, E.H., "Standards and New technology strategies to preserve content on magnetic and Disc Media, Point of View, SMTPE Journal, Vol.109, pp. 628-635, 2000.
- [5] Bearman, D., "Collecting Software: A new challenge for Archives & Museums", Technical Report 1, Archival Informatics, 1987.
- [6] Rothenberg, J. "An Experiment in Using Emulation to Preserve Digital Publications". Koninklijke Bibliotheek, Hague, Netherlands, 2000.
- [7] Granger, S., "Emulation as a Digital Preservation Strategy", *D-Lib Management*, Vol. 6 No. 10, 2000.
- [8] Russell, K., "Digital Preservation and the CEDARS Project Experience", Proceedings of the International Conference on Preservation and Long Term Accessibility of Digital Materials, York, England, pp. 139-154, 2000.
- [9] Woodyard, D., "Digital Preservation: The Australian Experience", Proceedings of the Third Conference on Digital Library: Position the Fountain of Knowledge, Malaysia, 2000.
- [10] Gilheany, S., "Preserving Information Forever and a Call for Emulators", Proceedings Digital Libraries Conference and Exhibition: The Digital Era: Implications, Challenges, and Issues, Singapore 1998.
- [11] Garrett, J, Waters D., "Preserving Digital Information", Report of the Task Force on Archiving Digital Information. The Commission on Preservation and Access and The Research Libraries Group, Washington DC and Mountain View CA, May 1996.
- [12] Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-R-1, Consultative committee for Space Data Systems, May 1999.
- [13] Shepard T., MacCarn, D. "The Universal Preservation Format: Background and Fundamentals". Proceedings of the Sixth Delos Workshop, Portugal, 1998.
- [14] Heminger, R.A., Robertson S.B., "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents", Proceedings of the Sixth Delos Workshop, Portugal, 1998.
- [15] Boudrez, F., "Archiving electronic office documents", 2003.
- [16] Abrams, S.L., Seaman D., "Towards a global digital format registry". In proceedings of the Information Technology and Preservation and Conservation Workshop in conjunction with World Library and Information Congress: 69th IFLA General Conference and Council, Berlin 2003.
- [17] Harvard University Library. "Library Digital Initiative", January 2003.
- [18] MIT Libraries. "DSpace: Durable Digital Depository". May 2003.
- [19] Public Records Office. "PRONOM System User Guide", November, 2002.

- [20] ISO TC 171/SC 2 N. "Document management — Electronic document file format for long-term preservation — Use of PDF (PDF/A)". September 2003. "
- [21] Dollar Consulting, "Archival Preservation of Smithsonian Web Resources: Strategies, Principles, And Best Practices", June 2001.
- [22] Besser, H. "Digital longevity". Chapter in Maxine Sitts (ed.) Handbook for Digital Projects: A Management Tool for Preservation and Access, Andover MA: Northeast Document Conservation Center, 2000, pp. 155-166.
- [23] Ockerbloom, J. "Mediating Among Diverse Data Formats", Doctoral thesis, January 1998.
- [24] Holdsworth, D., Wheatley, P. "Emulation, Preservation and Abstraction". Technical Report of CAMiLEON Project, University of Leeds.
- [25] Van Horik, R. "Image formats: practical experience". Contributions at ERPANET Training File Formats for Preservation, Vienna, May 2004.
- [26] Moehle, F. "The Role of File Formats in Digital Preservation: Opportunities and Threats". Contributions at ERPANET Training File Formats for Preservation, Vienna, May 2004.
- [27] InterPARES Project. Available online at <http://www.interpares.org>.
- [28] National Library of Australia. Available online at <http://www.nla.gov.au/>.
- [29] Clausen L.R. "Handling file formats". Report of the State and University Library, Denmark 2004.
- [30] "the top file extensions Windows site", URL: <http://www.icdatamaster.com>
- [31] "Wotsit's Format", URL: <http://www.wotsit.org>.
- [32] "My File Formats – the programmers file format collection", URL: <http://www.fileformats.com>.
- [33] "File Format Encyclopaedia", URL: <http://pipin.tmd.ns.ac.yu/extra/fileformat/>.
- [34] Internet Assigned Numbers Authority, "MIME Media Types", URL:<http://www.iana.org/assignments/mediatypes/>
- [35] Free Software Foundation, Inc., "GNU General Public License", URL:<http://www.gnu.org/copyleft/gpl.html>,1991.
- [36] Mellor P., Wheatley P and Sergeant D., "Migration on Request – a Practical technique for Preservation". In *Proceedings of the 6<sup>th</sup> European Conference on Research and Advanced Technology for Digital Libraries*, Springer-Verlag, June 2002.
- [37] Christensen S.S. "Archival data Format Requirements". Report of the Royal Library, Denmark 2004.
- [38] Miller D. R. "XML: Libraries' Strategic Opportunity". Net Connet Library Journal.
- [39] Digital Preservation Testbed White Paper "XML and digital preservation".
- [40] The Florida Center for Library Automation (FCLA). <http://www.fcla.edu/digitalArchive/index.htm>
- [41] OAIS blue-book, reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-B-1, January 2002, CCSDS. <http://www.ccsds.org/CCSDS/documents/650x0b1.pdf>
- [42] Hodge, G. Frangakis E., "Digital Preservation and permanent access to scientific information: the state of the practice". A report sponsored by The International Council for Scientific and Technical Information (ICSTI) and CENDI, February 2004.
- [43] Moore W. R., "Research and Development Plan and Schedule for the Research Project on Application of Distributed Object Computation Testbed

- Technologies to Archival Preservation and Access Requirements”. A report sponsored by National Archives and Records Administration and Advanced Research Project Agency/ITO.
- [44] John Mark Ockerbloom “TOM The Typed Object Model (TOM)”, document available on line: <http://tom.library.upenn.edu/>, 1999.
- [45] Ogbuji, U. “Thinking XML: The open office file format”, 2003. Available on line: <http://www-106.ibm.com/developerworks/xml/library/x-think15/>.
- [46] Barnette, A “Introduction to Computer Text File Formats”, 1994. Available on line: <http://ei.cs.vt.edu/~netinfo/notes/chap1/textformats.html>
- [47] Goethals, A. “Action Plan Background: WAVE”, 2004. Available on line: [www.fcla.edu/digitalArchive/pdfs/action\\_plan\\_bgrounds/wav.pdf](http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/wav.pdf)
- [48] The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage, “Materials Audio/Video Capture and Management”, 2003. Available on line: <http://www.nyu.edu/its/humanities/ninchguide/index.html>.
- [49] Verdegem, R. “Databases Preservation Issues”. Available on line: [http://www.digitaleduurzaamheid.nl/bibliotheek/docs/longterm\\_preservation\\_of\\_databases.pdf](http://www.digitaleduurzaamheid.nl/bibliotheek/docs/longterm_preservation_of_databases.pdf).
- [50] The National Institute of Standards and Technology (NIST), “Common Biometric Exchange File Format”. Available on line: <http://www.itl.nist.gov/div895/isis/bc/cbeff.old/CBEFFDraft2Cut031-REVNNov23.pdf>.
- [51] McGrath R.E., “XML and scientific file formats”, 2003. Available on line: [http://www.ncsa.uiuc.edu/NARA/XML\\_and\\_Binary.pdf](http://www.ncsa.uiuc.edu/NARA/XML_and_Binary.pdf)
- [52] National Initiative for a Networked Cultural Heritage (NINCH), “Audio/Video Capture and Management”, 2002. Available on line: <http://www.nyu.edu/its/humanities/ninchguide/VII/>.
- [53] Rauch C., Rauber A., ”A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies”, Report for the DELOS Network of Excellence, December 2004.
- [54] Chalmers R., “The Broadcast Wave Format– an introduction”, EBU Technical Review, Winter 1997. Available on line: [www.ebu.ch/trev\\_274-chalmers.pdf](http://www.ebu.ch/trev_274-chalmers.pdf).
- [55] Schüller D, “Digitisation -The Only Viable Way to Preserve Audio Recordings in the Long Term”. Contributions at ERPANET Training File Formats for Preservation, Vienna, May 2004.