Digitale Archivering in Vlaamse Instellingen en Diensten

DAVID

The digital recordkeeping system: inventory, information layers, and decision-making model as point of departure

Filip Boudrez

FACULTEIT RECHTSGELEERDHEID
INTERDISCIPLINAIR CENTRUM VOOR RECHT EN INFORMATICA
TIENSESTRAAT 41
B-3000 LEUVEN

Stadsarchief Stad Antwerpen

Version 1.0

Legal deposit library D/2001/9.213/7

Antwerp, June 2001

DAVID project website: http://www.antwerpen.be/david

E-mail address: david@stad.antwerpen.be

TABLE OF CONTENTS

I. INTRODUCTION

One of the first aims of the DAVID project[1] is to trace and list the digital archive documents in the custody of the Flemish administrative and records departments. Such an inventory would reveal the full spectrum of digital files and information systems held by the State and due for (digital) archiving in the foreseeable future. For a snapshot of the type of digital records the government currently keeps we examined the computer applications now running in the administrative departments of Antwerp city council, social security services, and port authorities.

The initial idea was to work out a typology from which a method for preserving the various types of digital archive documents over the long term would follow. This typology would stand or fall on its usefulness in formulating a preservation strategy, and was pursued with this goal in mind. The first attempt rested on the editorial form and function of the digitally preserved document[2], a method of description and classification borrowed from paper archiving. It was soon obvious, however, that this was no basis for managing digital archives and no basis for formulating a preservation strategy.

As things stand a number of general observations lead us away from the 'paper view' of digital archive documents. To start with, it assumes that digitally preserved documents are stored (separately) as documents, whereas the content of the digital file actually exists in the data stream of the information system. The (editorial) form we see on screen and on paper is ultimately defined by the structural information and computer application (see further). This form is not always fixed and can vary greatly. The same digital archive document can be stored as a text, spreadsheet or database file, or as a combination of the three. On top of this, many of our information systems[3] are designed to manage information without ever producing a document. And finally, digital information systems contain more than we see on the printed page. In a document-based approach much of the contextual information would be lost, whilst this is the very information needed to establish the authenticity of digital archives.

A digital preservation policy designed from the document perspective will soon veer to the problem of file format. Deciding a policy solely on this basis is not to be recommended since there is already much to identify migration as a long-term preservation strategy. The grave danger in this approach is that there may be little or no consideration of the information system in which the files are created. By linking our preservation strategy directly to the file format we lose sight of the architecture, operations and functionalities of the information system. We also ignore the possibility of interaction between digital files or information systems. By narrowing in on the format in which digital information is stored we run the serious risk of disregarding certain software, dynamic components, or external

---

1 Digital Archiving in Flemish Institutions and Administrations (DAVID) is financed by the Flanders Foundation for Scientific Research within the scope of the Max Wildiers Fund.

2 Providing proof, giving instruction, reporting, giving notice, assisting memory, planning and giving access (*Lexicon van Nederlandse archieftermen*, 1983).

3 "An information system constitutes the entirety of data, programming, procedures and equipment capable of transforming data into information" (P. HORSMAN, *Archiefsystemen en kwaliteit,* in *Naar een nieuw paradigma in de archivistiek*, page 88). In this report the term information system is intended as the entirety of data, structural information and application software used in a single system (see page 15). Information systems convert data to information. Document is intended here as paper information carrier. Digital archive components are indicated as digital archive documents.

information sources. Indeed, this type of dynamic and integrated information system has been on the increase in recent years. And there are several technical reasons for dismissing a decision model confined to file format in this way. As with applications, file formats can have numerous versions all differing from each other[4]. File formats frequently undergo internal alteration to make them compatible with the operating system or application that uses them[5]. Nor can we credit them with any real stability; they are far too vulnerable to commercial factors[6]. On account of this, the method gives us neither a systematic overview nor a stable basis from which to develop a preservation strategy[7].

With time it has become clear that digital archiving should be operated at a higher level, i.e. within the information system that creates the digital archive documents. We can classify information systems according to the file types they generate or the files that underlie them (text, spreadsheet, database and audio-visual). However, as with the paper approach to archiving, we can't work on the basis of system type without first looking into the properties (as described above) of the information system itself. Again, it is impossible in this sense to marry a preservation strategy to a single type of information system, particularly since the majority are made up of several types and possess a number of unique characteristics. This requires a thorough, preliminary analysis of each system and the solution will be system-specific in many cases. Therefore, we need to start from the information system itself. Of the information systems in use at the city council, social security services, and port authorities of Antwerp, a striking proportion store data in the form of mainframe databases and use ad hoc software. Until we have examined the full system we won't know whether it will suffice to preserve the mainframe databases as flat files, according to general convention. This reasoning also applies to simple computer applications such as ordinary text and spreadsheet files.

When it comes to working out a system for deciding preservation strategies we find that a typology of information systems actually brings us no closer. However, we do have a decision model that will

---

4 This is the case with file formats in Word97 and Word2000. Text files are perfectly interchangeable between the two versions, which leads us to assume that Word97 and Word2000 use the same file format. However, there is a very fundamental difference in the way each version composes its files. We see evidence of this in the different file sizes when the same text is saved as a Word97 or Word2000 file. Another well-known illustration of this problem is the development of graphic GIF files. We can not tell if we are dealing with a GIF file with animation (GIF89a), or without animation (GIF87), unless we view or open it with a web browser or appropriate viewer. In most cases file size is also a good indicator (GIF87 < GIF89a).

5 TIFF files are frequently altered internally to suit one specific software application. This gives rise to a huge variety of TIFF files that can only be displayed on screen in one specific application.

6 The use of the Graphics Interchange Format (GIF) is a fine example of this. In 1987 CompuServe provided GIF as an open and free specification. In the (de)compression of graphic data the GIF format uses the Lempel-Ziv and Welch algorithm. This algorithm stores images in small files enabling the GIF files to be quickly and easily exchanged. GIF became a standard. CompuServe was mistaken in its view that the LZW algorithm belonged to the public domain. The problems started when Unisys exercised its patent rights on the LZW (de)compression algorithm, meaning that all users now pay Unisys a license fee until the end of 2003. This implies that every records department that migrates images to GIF files should in theory apply for a licence. As a result of the general dissatisfaction with this course of events PNG was created (Portable Network Graphics; set up by W3C) as the successor to GIF, and this format is free of patent rights.

7 Over the years several investigations have centred on the difficulties associated with archiving certain types of file formats. See: G.W. LAWRENCE, W.R. KEHOE, O.Y. RIEGER, W.H. WALTERS and A.R. KENNEY, *Risk Management of Digital Information: a file format investigation*, Washington, 2000; J.C. BENNETT, *A framework of data types and formats, and issues affecting the long term preservation of digital material*, Wetherby; 1997. These investigations merely listed the basic points to bear in mind when digitally preserving several types of file format. In *Comparison of Methods & Costs of Digital Preservation* consultant T. Hendley takes the matter a step further. He differentiates 10 digital sources and indicates for each source type the file formats to which they should be migrated. He recommends emulation only for multimedia applications.

allow us to identify the most suitable preservation strategy, starting from the information system itself. This model rests on a recurring pattern of factors worth examining for every information system: WHAT do we archive? WHO manages the digital archive? HOW do we preserve the digital archive? WHEN does the transfer take place?

This report makes the case for a decision model based on the four questions above[8]. Each question is followed by a brief description of the factors to be accounted for. The result is a decision model that we can use in our systematic investigation of information systems. There is nothing random in the order of the questions. The question WHAT do we archive is best asked from the perspective of the information system as a composite of several information layers (see figure 2, page 15). Seeing which layers are preserved on the long term will, to a large extent, indicate the direction of the preservation strategy. The question of WHO manages the digital archive is closely related to this. The questions HOW and WHEN relate more to the practical side of the process. This decision model is explained in the second part of the report and provides the springboard for a digital preservation system. By way of introduction the report gives a brief explanation of the three preservation strategies used with digital archive documents.

Though the model covers a few indirect elements, such as recordness, selection, metadata, description and rendering, it does address the main challenge of the archiving system, i.e. to manage digital archive documents and preserve their readability. When it comes to preserving the readability and authenticity of digital archive documents, it is not enough to preserve the computer files alone. We need to preserve the emulations, metadata, log files, scripts, context, procedure descriptions, etc. too. However, this lies beyond the scope of the present report and will be dealt with in a later DAVID report on authenticity.

[8] With thanks to Inge Schoups and Willem Vanneste for their corrections and suggestions.

II. PRESERVATION STRATEGIES

The long-term preservation of digital archive documents actually constitutes a problem because the hardware and software environments in which the documents exist become outdated and obsolete. One quality peculiar to digital information is that it requires IT technology to be viewed, and IT technology is constantly evolving. If they do nothing, archivists and records managers will be faced with digital archive documents that they can no longer access with more recent IT technology. In broad terms there are three preservation strategies for digital archive documents[9]: hard copy, migration, and technology preservation (including computer museums and emulation). It is generally agreed that hard copy and computer museums should be set aside for emergencies, leaving us with migration and emulation as the only real options for the time being. Having said that, there is no real consensus over the usefulness and advantages of migration and emulation. Migration is about altering digitally preserved files to suit a new technology; emulation is about mimicking the original technology in a new environment and then consulting the archived file in its original format.

II. 1 HARD COPY

In the hard copy strategy the digital information is simply printed or put on microfilm. The only advantage is that the information is now stored on a durable medium and we can safely ignore the thorny problem of hardware and software support. The obvious disadvantage is that we lose all the functionality and all the advantages of digital information, and not all multimedia files can be stored in this way. For example we can't store moving images or sound. It is also difficult to guarantee authenticity with hard copy because much of the original properties are lost (e.g. form, structure, context...). There is also more storage space needed than for digital files. For that matter, a study by the NARA has shown that digital preservation is less expensive than printing the digital information on paper[10].

---

[9] K. THIBOUDEAU, *The unsteady state of the art of preserving electronic records*, Lecture given at the VIth European Archive Congress in Firenze on 31 May 2001.
Some authors see the use of standards as a fourth strategy. In practice this route is closely related to migrating digital information. However the life of the standard is also limited in time. In small applications standardisation is mainly a question of file format. In large information systems (such as mainframes) the type of database application is important.

[10] J. HOFMAN, *Het 'papieren' tijdperk voorbij. Beleid voor een digitaal geheugen van onze samenleving*, The Hague, 1995, p. 24.

II. 2 MIGRATION

As a strategy for preservation, migration generally involves transferring files to another environment so that they are compatible with a new or different computer configuration. In its broader meaning migration relates to the hardware, operating system and application software. In other words, the files are adapted to the new environment in which they will be used in future. Migration can sometimes mean refreshing (transferring to another medium) or conversion (transferring to another version of the same application). In its narrower sense migration means conversion to a different operating system[11].

When transferring digitally preserved information the choice of file format is crucial. The files are converted to a format that is more durable, or better suited to long-term preservation. The preference lies, where possible, with a standard file format and there are a number of factors in support of this. The first is that the software manufacturers implement standards and abide by them when developing their applications. This makes the interchange of computer files possible. Secondly, when stored in a standard file format digital information can be interpreted by several different applications. This reduces the chances of a format being migrated if one of the applications becomes obsolete or is no longer supported. A standard file format is more durable and usable than a non-standard format. But beware. Software manufacturers often add application-linked properties to the standards in order to consolidate or increase their share of the market. Standards have an equally finite lifespan and will themselves be migrated in future, albeit less frequently than non-standardised formats. Finally, a standard's usability depends on its application in practice. We can only speak of standardisation if manufacturers and users apply a standard liberally.

File formats fit into a hierarchical structure. At the top there are the official standards. These are set by the *official* standardising authorities, such as the *ISO (International Standard Organisation*), which may work in co-operation with the *IEC (International Electrotechnical Commission)* or the *ITU* (*International Telecommunications Union)*. It is no coincidence that the formats used to exchange data constitute the largest group of official standards. The most widely recognised examples for text files are ASCII, Unicode and SGML. The description of *ISO* standards are not free of charge and not all official standards are widespread.

Next in the hierarchy, after the official standards, come the *de facto* standards. Although the term 'standard' is actually reserved for (inter) governmental organisations (*ISO*, *CEN*, *ETSI*, *ANSI*, etc.) the de facto standards cover file formats that become normative as a result of their widespread use. The de facto standards fall into three groups: specifications or recommendations by the standardising organisations, open file formats dependent on a single manufacturer, and closed file formats.

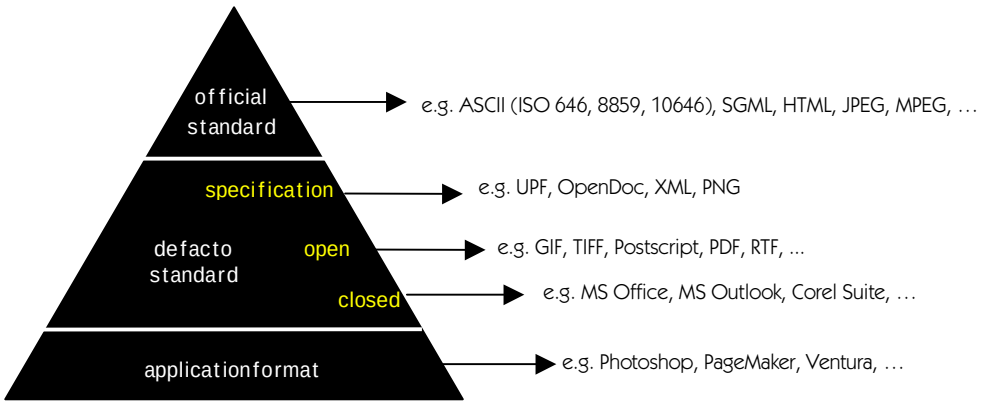
The specifications or recommendations are the result of joint venture initiatives aimed at putting norms and standards in place. One of the best-known current examples is *W3C* (*World Wide Web Consortium*). As with the official standards the procedure involved in writing up the specifications for these file formats is very long and complicated. This, along with the fact that several parties are

[11] Conversion and migration are frequently used to mean the same thing and there is no real consensus as yet as to how the concepts should be distinguished. Conversion is sometimes used to describe the transfer of data to another software application (e.g. Lotus → Excel) and migration to describe the transfer of data to another platform (e.g. Unix → WinNT). (*Begrippenlijst digitale duurzaamheid; Ontwerp-Regeling geordende en toegankelijke staat archiefbescheiden 2000:* art. 1; C. DOLLAR, *Authentic Electronic Records: Strategies for Long-Term Access*, p. 29-31;). Others use conversion strictly to mean the preservation of data in another version of file format and migration to mean transfer to another file format (G.-J. VAN BUSSEL, *De opmaat voor hoe het niet moet*, in *Archos Magazine*, 11 (2000), page 5).

involved in the initiatives (software manufacturers, universities, consumers) goes some way towards guaranteeing stability. We cannot underestimate the effect of these specifications. From a strictly legal perspective the official standards are more highly thought of, but there is a greater chance, thanks to the interest and involvement of the large software manufacturers, that a number of these specifications will find a wider application than the official standard (XML versus SGML for example).

One element shared by open and closed file formats is that they belong to a single manufacturer. They are different in that with open formats the specification is released and with closed formats it is not. At the very bottom of the hierarchy we have the file formats used in more obscure commercial or ad hoc applications. These formats are not easily interchangeable and are dependent on one application.

Figure 1: Hierarchy of file formats



The difficulty with migration is that suitable formats are not always to hand. Archivists can only use the formats available on the market and this is dependent on private initiatives and standardisation projects. In many cases migration means loss (e.g. layout, functions, etc.) and so priorities have to be set. This comes down to deciding the most important elements of the digital archives beforehand (e.g. content, structure and context), then preserving them in the new format. This was the method used to digitally archive the electoral register[12].

Migrating archived files to a suitable preservation format will seldom be a one-off operation. There is a particularly high chance of repeated migration when using open and closed de facto formats and application formats. Every migration or conversion must be documented (original file format, method, test, loss of information, etc).

[12] The DAVID report on the digital archiving of the electoral register is available at the DAVID website (under publications → reports).

II. 3 TECHNOLOGY PRESERVATION

If file formats are not adapted to the new technology then the original technology has to be preserved or simulated in some way or other. There are several technical options here, such as preserving the original hardware and software or using configurable chips and virtual machines. Emulation is probably the best-known example of putting obsolete technology to use in the future.

Computer museums employ a strategy of preserving the original hardware and software required to consult the archived files. The advantage with this is that the archive documents and the hardware and software they require are preserved in authentic and original form. The archives remain available in digital form and there is no concern with migration or emulation. For the most part, management and maintenance is confined to the regular transfer of files to other carriers. However the costs and expertise required, the ephemeral nature of the disks and chips, and the lack of commercial support make this solution less likely, if not unrealistic.

Emulation is a strategy in which files are stored in their original format and the requisite hardware and software environment is mimicked on a "host" system[13]. Emulation can be applied to hardware, software, or both. One of the simplest applications is hardware and operating system emulation, and in this case the strategy involves the following steps:

- ✓ Writing an emulation specification for each platform. This specification contains information on the hardware and software requirements and includes the register, jumper settings, operating commands, file formats, etc. The specification is written in a specific language (*emulator specification language*).
- ✓ Creating an interpreter suited to the emulation specification and adapted to the virtual machine. The interpreter is a program that runs on a virtual machine (VM) and creates the actual emulation of the old platform. A new interpreter is needed for every specification language. An interpreter is a sort of compiler that converts readable source code to machine language.
- ✓ Starting the original application on the emulator.
- ✓ Retrieving the archived digital files from their (original) application. The original digital file and the original application, or the software used to render the document, must be archived with the specification and interpreter.
- ✓ Once a VM becomes obsolete the specification and interpreter have to be written for a new VM capable of simulating the old. The specification and interpreter for the old emulator can then be run on the old VM. The old VM runs on the new VM. The archived digital file can then be rendered as follows: 1. Start the interpreter for the corresponding VM on the current VM, 2. Start the interpreter for the appropriate application, 3. Open the archived file.

With this type of emulation application we need to preserve the application software and the emulation specifications along with the archived files. Metadata are very important in emulation. In addition to the (obligatory) metadata we need to preserve the emulation specification. Since this information should be to hand at all times, it is recommended that the specification be printed on

[13] Http://129.11.152.25/CAMiLEON/dh/ep5.html; J. ROTHENBERG, *Avoiding technological quicksand*, 1998; J. ROTHENBERG, *An experiment in using emulation to preserve digital publications*, April 2000; S. GILHEANY, *Preservation Information Forever and a Call for Emulators*, Singapore, 1998.

paper. The key to emulation is ensuring that the basic functionalities of the information system are retained. In theory there is no need to emulate extra or non-essential functions. A minimum version of the software should do in most cases. Specific knowledge could still be needed to operate the emulation, and so a users' manual will be required too.

The most widespread development environment for emulations is found in C/C++ programming languages. Emulation is already the general method of keeping video and computer games operational. As yet there is no real consensus as to when emulation software should be created. Rothenberg maintains that it will keep until the platform becomes obsolete, whereas others, such as Holdsworth and Wheatley, find it important to start developing emulation programs while the original applications are running.

III. DECISION MODEL FOR A PRESERVATION STRATEGY

Both digital preservation strategies have their protagonists and antagonists. The advocates of emulation (e.g. J. Rothenberg and S. Gilheany) argue that emulation is cheaper and less labour intensive. Keeping files in their original format retains their functionality and form, and their 'look and feel'. These elements constitute a measure of authenticity and integrity, and, in many cases, are (partly) lost with migration. The opponents of emulation point out the need to migrate the VM's, and question the technical feasibility of emulation in general. The initial results of emulation tests seem to back them up in this, but more study is needed before judging the emulation issue once and for all. The advocates of migration (such as C.M. Dollar, D. Bearman,) are well aware of the weaknesses of the migration strategy and admit that there are situations in which it is not technically feasible to preserve functionality and integrity with migration. In these cases the archivists need to consider another route (emulation, hard copy, etc.) or prioritise and archive only the essential data[14].

The ideal situation from an archival point of view would be to cut hardware and software dependence to a minimum and involve software as little as possible in the preservation process. The migration route, in which files are transferred to a standard format, comes closer to this ideal than emulation, but judging from our inventory of digital information systems at Antwerp City Council, it is not expedient to make a policy decision on 'migration versus emulation' as a preservation strategy. In practice we need to remember that every bit stream relies on the appropriate software to give it significance. Software dependence varies from information system to information system. The crucial thing is whether the digital files can be archived independently of the software environment in which they were created. Emulation may be recommendable for one information system whereas migration will suit another. For example, migration is the ideal route for preserving the electoral and population registers. GIS (Geographic Information Systems) -applications are so strongly integrated and so software dependent that emulation would appear best for them. Dynamic web pages in ASP or JAVAscript can only be viewed through a web browser with the appropriate webserver and/or -client software. No doubt file dependency and the future functionality of the archived files will be determining factors in the choice of strategies. After all, digital files can be dependent on specific software applications and external files. Nor can we, for that matter, rule out a consecutive migration and emulation phase (or vice-versa), in the lifecycle of the digitally preserved file.

In general we can say that migration is the best solution if all we want to do is archive the data stream and confine functionality to the retrieval and consultation of fixed information. Emulation is best if, for the sake of functionality, we also want to preserve the structural layer or tools. In any case migration has the edge over emulation in that it is technically less complex and something that records departments could do for themselves. Emulation, on the other hand, is the work of specialists and involves preserving emulations (specifications, interpreters and VM's) alongside the archived files.

---

[14] C.M. DOLLAR, *Authentic Electronic Records: Strategies for Long-Term Access*, Chicago, 2000, page 72-74; D. BEARMAN, *Reality and chimeras in the preservation of electronic records*, in *D-Lib Magazine*, April 1999.

III. 1 INVENTORY OF DIGITAL INFORMATION SYSTEMS

When deciding a preservation strategy, as we have seen, the best place to start is with the information system itself. Obviously, the archivist needs to know exactly which records the administrative department is creating. A systematic summary, such as an inventory of the various digital information systems, is important for several reasons. It is necessary for the archivist to be aware of all the information systems run by the archive-creator, and the properties of the IT-applications they require. This tells him what information is held in digital form at what location. He should be able to determine the archive value of the information from the data kept on each information system. This should also enable him to formulate a policy for preserving digital information of archive value. An inventory of this type can serve as a meta-information system, from which the necessary metadata can be taken at a later date. One of the important elements that must be described is the context of the records and the relationship with other paper or digital records. This data are also important to prove the authenticity of the records. The inventory is one of the few places where these information can be registered.

While gathering data on the information systems of the city council, social security services, and port authorities of Antwerp, we noted that they were predominantly designed to automate working procedures and operations, and so a functional inventory was called for. Setting the context in this way and briefly describing the functions of the system gives the archivist a basis for ascertaining the archive value. An up-to-date inventory of the administrative department's digital information systems is actually a key instrument in the digital preservation policy of any records department. The inventory is not included in this report, but is a separate working document requiring constant monitoring[15].

Compiling and maintaining an inventory of information systems is no easy task. Experience in the Netherlands and at home reveals that administrative or computer departments seldom keep the basic information needed in systematic form. From this observation has grown the idea of bringing in a new rule that an inventory of digital information systems should be kept[16]. A further difficulty encountered when gathering information on an organisation's digital archives is the matter of hard disks on local workstations. In many cases digital archive documents are kept on C- and D- drives with no effective controls or management practices. An inventory may draw the administration's attention to this problem and make them more aware for the issue of digital record keeping.

[15] The inventory of digital information systems can be viewed at the DAVID site (under 'publications' and 'other publications').

[16] An inventory such as this, for example, is prescribed in the *Besluit Informatiebeheer Provincie Zeeland 1997*, art. 12: "The head of the custodial body shall ensure that a list of the information files is compiled and maintained, in which the information files are described and linked to the various work processes and tasks". This obligation is also included in the *Ontwerp-Regeling geordende en toegankelijke archiefbescheiden 2000* (article 11) and is a milestone in *the Modernising Government 2004 Requirement* of the Public Record Office.

III. 2 INFORMATION LAYERS AS THE STARTING POINT

One thing that prevents us from laying down a typology as the basis for a preservation strategy is the huge diversity of information systems as a whole. The file formats in which the digital information is stored, the configurations, the software applications used, the nature of the information, the functionalities, and the dependencies differ from system to system. Nonetheless, there are a number of common elements that may shape our decision as to which preservation strategy to pursue:

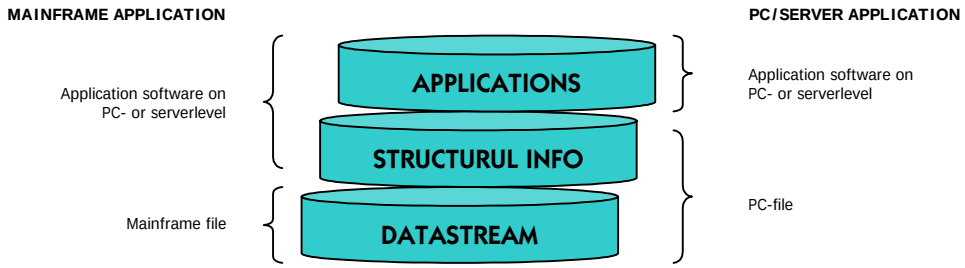
- ✓ The information systems are designed to automate the functions of government. Special information systems have been developed for the major or important functions. The files that flow from one functional work process (text, database, spreadsheet) are generated in the same system and linked to each other in many cases. Information systems too may be linked to each other and cannot therefore always be preserved as isolated entities. This has led to the conviction that the information systems themselves should be the starting point, not the type. Since we observe an obvious relationship between the information systems and the functions of the creator, it is easier to classify our overview of the information systems in terms of function. This link with the functions is an important as a way of setting the context and establishing archive value.
- ✓ Digital data are stored in files. The files are constructed according to a code and arrangement peculiar to that application. Every format will code, arrange and order the bits in a unique way. In general the arrangement breaks down into two components: the structural or logical layer on the one hand, and the data stream on the other. The structural or logical elements constitute information on data used by application when performing its functions, and are therefore peculiar to application and file format. The data stream is the series of bits containing imported or generated data. This part of the bit stream can signify anything: text, sound, image, video, etc.[17]. Without structural information the data stream cannot be (correctly) rendered, with the exception of flat files, provided the bits are reproduced as ASCII characters[18].
- ✓ One factor linking text files, spreadsheets and data banks on PC's and servers is that the bits in their data streams represent ASCII or Unicode characters. Their structural elements differ because they depend on application type, application used (product) and version. The data stream is the core of every file or information system, and the first thing that should be preserved. Whether the structural elements should be archived too will depend on the future functionality of the archived data elements and the operations still to be carried out.

[17] By way of illustration: one structural element of a spreadsheet or database is the header, which contains information on the respective cell arrangement and formula or fields (name, length, data type, etc.) and index. A word processing file of the type found in MS Word is an Object Linking and Embedding application, which generates multistream files. These stream files are linked and form a single file. An MS Word97 file consists of a main stream (including header, formatting information), summary information stream, table stream (including plcf's), data stream and 0 or more object streams. The four bit streams alongside the data stream are structural elements. A GIF file consists of a header, a global and a local colour table (both optional), a local image descriptor and the image data. PDF files are internally structured as follows: header, body, cross-reference table, and trailer.

[18] In the OAIS model the information system consists of five layers: physical, binary, structure, object and application. The three layers we describe correspond to structure, object and application. Though the physical and binary layers are important we have omitted them because in theory the choice of medium (physical layer) is free, and the file system (binary layer) is related to the operating system.

Whenever digital information is compressed for storage we must ensure the appropriate tools for decompression, or it won't be possible to view the information later. In this way we can split every information system into layers. Our examination of the system should then reveal which layers are to be archived with the data stream. Where no special functionalities exist and no dynamic components are archived, the means of digital preservation points towards the data stream. In the other case, structural information and tools must be preserved alongside it.

Figure 2: The layers of an information system.



✓ Most information systems contain only textual information kept in the form of databases. This data stream can be read by several applications. Of the 59 information systems belonging to the administrative and social security services of Antwerp, whose application type could be discerned, databases (86%) represented the core application for 51. Databases are generally the most frequent computer applications, and the largest group of information systems of archive value[19]. If it will suffice to preserve only the data stream, the files can be archived as flat files. In the case of PC/server applications the data stream is filtered, as it were, and stripped of the structural layer. The result is an ordinary ASCII or SGML/XML file. A file transfer is performed for mainframe files, in which the EBCDIC characters are converted to ASCII.

EBCDIC (Extended Binary Coded Decimal Interchange Code) is IBM's 8-bit expansion of the 4-bit Binary Coded Decimal code table and forms the basis for the code tables used by mainframes to this day. There are a number of code tables. The code tables for Belgium are 274 (Belgian) or 500 (Belgian New). It is also possible to set up your own code table. In Antwerp, a code table based on no. 274 is used with BEAM, but it contains extra characters to permit a number of foreign symbols. The mainframe code table is then controlled from the application. Mainframe files can be transferred to server or PC level via ordinary FTP or SNA transfer. There are even a number of specific tools available for this. The EBCDIC code is converted to ASCII in the process. This conversion is almost never error-free. Most of the problems relate to the conversion of diacritical symbols. For a file transfer from EBCDIC to ASCII we need a COBOL copybook (*.LAY: data lay-outfile) describing the record layout (structure, field length, etc.) of the files. The final result of a transfer of this type is an ordinary flat file.

[19] The huge proportion of database applications in the world of information systems is confirmed by data on the nature of computer applications at the Dutch and Canadian governments (A.A.C. JANSEN, *MLG's geteld en gewogen*, in J. HOFMAN (Ed.), *Het papieren tijdperk voorbij. Beleid voor een digitaal geheugen van onze samenleving*, page 64, 85; G. BLAIS, *L'expérience des Archives nationales du Canada*, lecture at *Journées internationales: La conservation à long terme des documents éléctroniques*, *Paris 8 March 2001*).

When converting mainframe files to the server or PC level the file size is as important as the diacritical symbols. Mainframe files are always too large to be used as one server or PC file. In most cases they will have to be split into smaller files.

III. 3. DECISION MODEL

Separating the layers of an information system is a good basis for examining the best method of preserving a particular information system. In mainframe applications the layers are more clearly separated than in ordinary PC applications. For that matter, in future the distinction between the data layer and the structural layer should be much clearer in ordinary PC applications. The use of SQL databases and the SGML/XML file format is a fine example of this development. The first question relates to this.

→ What do we archive: only the data stream? The data stream combined with the structural layer? The application tool? Can we select? Will we transfer different versions?

→ Who manages the digital archive: the creator, the computer department or the archival service?

→ How do we preserve the digital archive: in which file format? In what way does the records or archival service take possession of the archive?

→ When is the digital archive transferred: at the end of the administrative storage term? After an upgrade? After reaching a certain file size? What periodicity?

The answers to these questions will help us form the main lines of our preservation strategy. Together they are the cornerstones of a model that will help us decide how the digital files should be archived. For the purpose of illustration, reference is occasionally made to the information systems at the administrative departments of the city council, social security services or port authorities. For more information on these systems please refer to the *Inventory of digital information systems* (see DAVID website).

III. 3.1 WHAT do we archive?

Although in each of these questions we start from the information system itself, this does not mean that we should preserve the entire system and all of its layers. This may be the case for some information systems, but for others it will be enough just to archive the data, or part of it at any rate. In practice other information will need to be stored alongside the archive data, but here we focus only on the digital documents to be preserved.

As a rule paper documents go through a process of selection before transfer. The files to be archived are separated from those elements that can be destroyed. The choice between preservation and destruction is largely determined by the legal or historical import of the documents. In a digital environment this choice depends not only on the archive value, but also on the question of whether it is technically feasible to destroy the digital data. Separate or independent files (such as word processing files) can be destroyed without problem and without disrupting the information system.

It becomes more difficult when we need to permanently preserve only some of the data in large or integrated information systems. Will the part of the file to be permanently preserved remain after the rest has been destroyed? Will the application still function without the destroyed information? Will it still be possible to consult the permanent archives with ease? What about the integrity of the file? A good example, perhaps, is FIN2000, a mainframe application used to store the council's accounts. FIN2000 contains documents which have to be archived permanently (annual accounts, ledger, etc.) and temporarily (daybook). Can the daybook be destroyed without affecting the accessibility of the annual accounts or the application's overall functioning?

Thus, with digital data, not only do we need to check which digital files are of archive value, but we also need to examine the technical feasibility of separating the information we wish to preserve from the information earmarked for destruction. In most cases it just won't be possible to archive that part of a file or information system that is worth archiving and destroy the rest. In these cases the choice will be confined to preserving or destroying the whole of the file or system. In other words, selection will no longer take place at the digital archive level, but at the information system level. In the future it may well be the case that information destroyed in the paper environment is preserved digitally nonetheless, because it is inextricably linked with the data for permanent preservation or with the information system itself. This implies that selection lists compiled for paper documents are not necessarily applicable to digital archive documents.

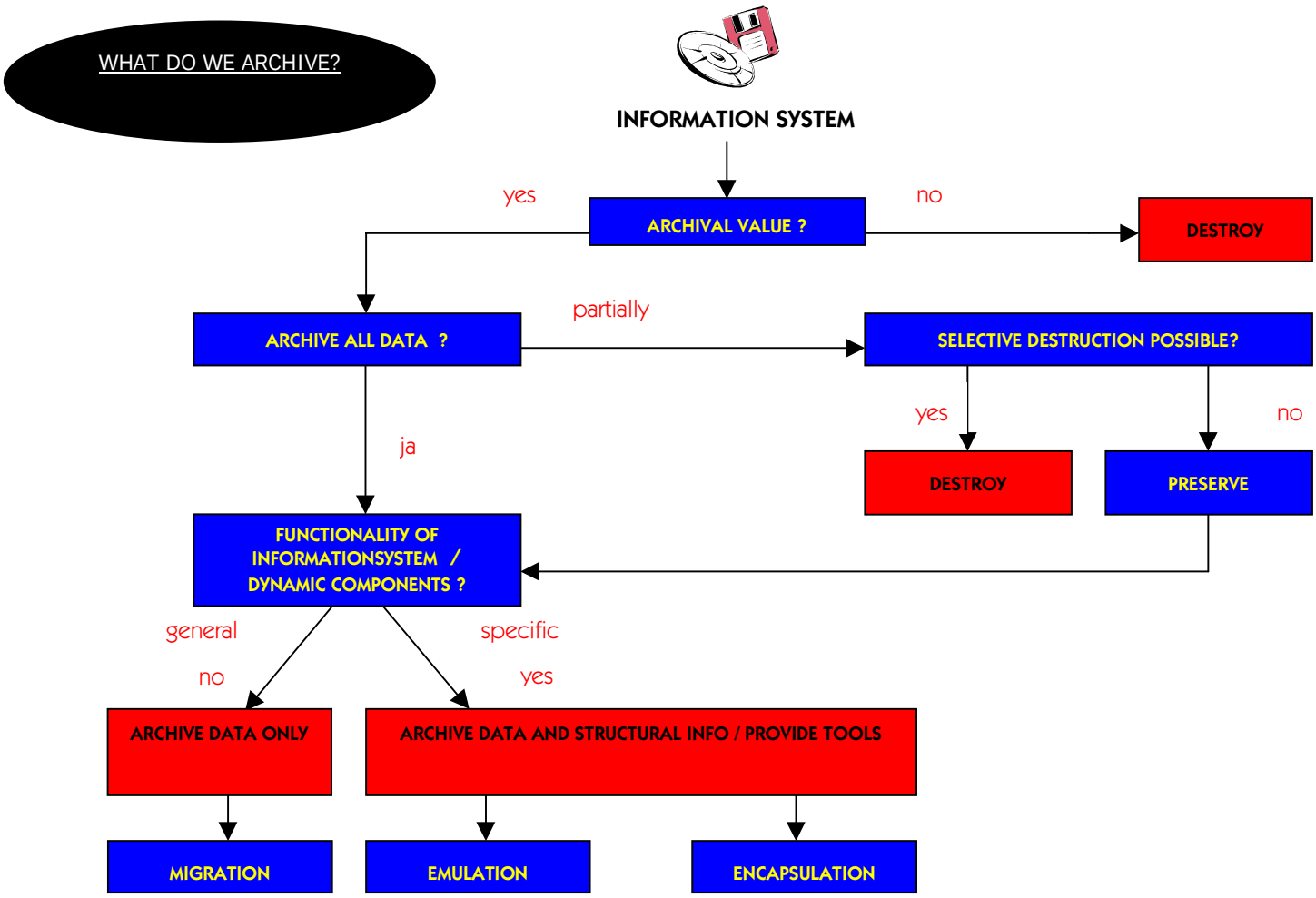
Another approach may involve exporting that part of the digital information to be preserved from the system and then archiving it separately. In taking this route we will be preserving the data outside of the original information system and will need to establish elements such as context in another way, for the purpose of authenticity for example. We are currently in the process of examining the extra data to be archived along with the digital data.

Secondly there is the question of whether it is enough to archive the data and structural information of an application, or whether we should archive special tools or software as well. Digital information is characterised by the fact that it can only be consulted with the help of technology. This is superfluous for data with general and widespread functionality, such as text or ordinary data files, in which the tool functions merely to retrieve and consult the information, and it will suffice just to archive the data. In the future there will be plenty of applications to support these basic functions. One point that came across in our description of system types is that many computer systems use specially designed and custom-programmed tools. One example is the GIS viewer, KAVIA, developed by Telepolis. Can the information from this application (e.g. land register information linked to the GIS map) still be viewed, retrieved or compiled without the KAVIA viewer, or by other, imported viewing tools? If not, the specific functionality of the application, and the software required, must also be preserved. For that matter, we have seen the digital files of specific software packages become increasingly dependent in recent years. This is a development that goes hand in hand with the ever-greater spread of multimedia applications, and the ongoing integration of information systems as a whole. This is why preserving an information system like GIS causes so many problems. The GIS components are strongly interwoven and dependent on specific software applications. The

preservation of compressed files poses a similar problem. We have to decide, judging from the functionality of the files or information system, whether to archive just the files containing the information or (a part of) the information system. Preserving (a part of) the information system implies storing software along with the data files. In a number of cases this may even mean preserving the original application. When specific tools are archived with the data we need to ensure that the tool can still be used and that the requisite hardware configuration and appropriate operating system are available. Sooner or later this route will require an emulator of the platform. Much will also depend on the way in which the digital information is made accessible.

If we transfer digital information to the records department together with an application, we should remember that a software license might be required.

Furthermore, the records department needs to keep metadata on the archived digital information and any other extra tools. These metadata generally relate to the technical properties and contextual data of the archived digital information. The metadata may also include the emulation specification. Metadata must be readable at all times. Consequently, they are preserved as ASCII files or stored on paper. It is preferable that metadata are standardised. They are provided, in the majority of cases, by the creator and the computer department.



III. 3.2 WHO manages the digital archive?

When paper archive documents are thought to be of archive value they are transferred to the archival service on expiry of their administrative storage term. Thereafter, the archival service has full custody of the archives. With digital archive files it is not necessarily the case that documents for permanent preservation will be transferred to the archival service. In a digital environment the archival service, computer department, creator, or even all three, may have custody of the archives together.

Non-custodial

The Australian archive departments adhere to the *non-custodial* view. It entrusts custody of the digital archives to the agency that created them. After all, these archive-creators have the technical infrastructure and knowledge needed to manage the digital systems. For the most part this concept reduces the tasks of the archive department to selection and exclusion. The Australian concept takes the view that it is not feasible for the archive department to have at its disposal every platform used by the administration. This view accords with its opinion that the cheapest and best solution to digital preservation is custody for as long as the original hardware and software configuration will allow. As long as the system operates there is no need to migrate the digital files. However, the application must be capable of managing an accessing the archived files. Most systems do not yet meet this requirement, so when using this method of preservation it is all the more important for the archivist to supervise the system's development. Through this approach the Australian archivists hope to skip a number of migration steps, which should bring cost and labour savings. Moreover, the accessibility and authenticity of the files are assured because they remain in the information system in which they were created[20]. This type of solution can prove useful for applications that are kept on specific platforms (e.g. mainframe, Unix, etc.) or require special applications which are not at the archive department's disposal. Antwerp city council keeps lots of large files at the mainframe level, none of which can be placed at the archival department's disposal.

Custody in the original environment can last only while the software environment is operational or fully supported. When this is no longer the case the time for decision arrives. Either the archived files remain in the custody of the department that created them, and are transferred to a new version or application where they can be managed and consulted via the active application, or, they are transferred to the archival service and then migrated to a format it supports, or they are kept in an emulation of the original system. One difficulty that arises with a *non-custodial* preservation policy is the availability of archived data to researchers and citizens. Researchers need to apply to the administration for access to the digital archives, unless the archival service provides access to the information system.

In object-oriented information systems this concept is close to being applied. With an application of this type a history of every object can be kept. In this way the original system can manage 'old' information or information 'for preservation'.

[20] Http://www.naa.gov.au/recordkeeping/er/keeping_er/; *Managing electronic records issues. A discussion paper*, April 1998, page 31-32.

Custodial

The American and Canadian Archives use the *custodial* concept, meaning that they themselves are responsible for managing the digital archives. The basic requirement here is that the archive department can provide the technical support needed. Much depends on the computer facilities available to the archive department and the way in which the digital depot is organised.

Before the digital files arrive the archive department must be sure that it can provide the hardware configuration, operating system and application software required by the files at its disposal. There is a limit, however, to the technical support an archive department can provide. If there were not, the archive department would be in danger of becoming of a computer museum. Yet in practice we know that administrations have information systems with a huge diversity of technical requirements in the areas of application type (mainframe, server, PC), platform type (mainframe, WINNT, Unix, Novel, etc.), and ad hoc application software. The upshot being that files for transferral must be altered to comply with standards at the archive department (migration). These standards can apply to the platform, file format, code tables and even media, and are best recorded on a list of archiving standards. After the inventory of digital information systems this list is a second important policy document. It should be regularly updated in line with the latest IT developments and contain the quality requirements to be met by the digital archive documents on transfer[21]. Another option is that the archive department alters its infrastructure so as to support the original files (emulation). This enables the simulation of hardware configurations, operating systems and even applications that would not otherwise be present in the archive department's infrastructure.

Obviously, it is easier for the archival service if the agency creating the records does not use too many information systems with widely varying technical requirements. When it comes to digital archives a little standardisation in this area would simplify the archivist's task. If he is involved in the development of a new information system, he should supervise upon that. Ideally, of course, standards should be implemented in the information systems, so that files for preservation can be archived with a minimum of changes.

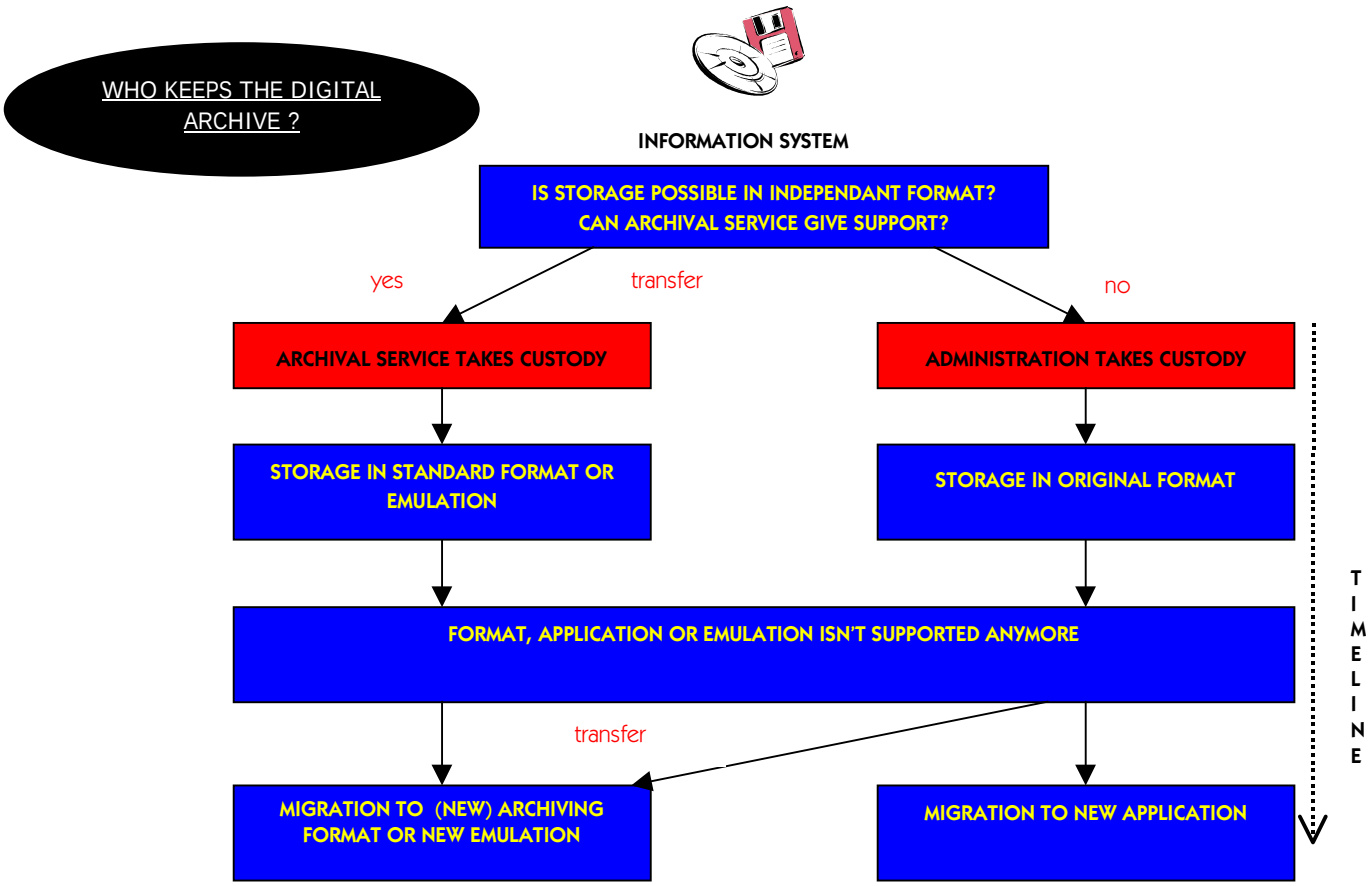
Digital preservation also means checking the method of transfer. The archive documents can be transferred to a portable medium (disk, CD-ROM, DVD, tape, etc.). In this case the records department must have the apparatus and drivers needed to read in the files. Another possibility is that digital files can be transferred over a network. This avoids the problem of having to read the portable medium, but could pose problems for very large files. The problem of building a digital depot is closely related to this. The records department can build a physically separate digital depot comparable with a separate domain. The file servers and jukeboxes would need to have the necessary storage space. More than 90% of today's set-ups are DAS (Direct Attached Storage) -applications in which disks and RAID systems are directly linked to servers and clients. Thought on the future of data storage is moving towards the development of SAN- and NAS-applications. In the SAN (Storage Area Networks) configuration the disks and RAID systems are linked via a network to the servers. In the NAS set-up the storage devices and server form a whole (Network Attached Storage). In the digital preservation of documents it seems that the NAS set-up offers added value, because heterogeneous platforms can exchange and store data on the same NAS server[22]. In NAS and SAN-like

---

[21] Examples of lists of this type: http://www.naa.gov.au/recordkeeping/er/keeping_er/append_a.html or annex 1 to the *Ontwerp-Regeling geordende en toegankelijke staat archiefbescheiden 2000*.

[22] G.-J. VAN BUSSEL, *Toekomst in opslag: NAS en SAN*, in: *Archos Magazine*, 7/8, 2000, page 4; *Software Development Network & Storage Architecture Guide* and *Storage Architecture Guide*

configurations no physical storage place for files used by the administration and archived files is distinguished. Here, transfer to the records department relates to the user rights granted, or can be performed by placing the archived digital files behind the archival department's firewall. With this type of digital depot there are no worries about the media on which the digital files are placed.



(http://www.auspex.com). According to Auspex, the producer of NAS servers and technology, NAS applications have the edge over SAN because they support the file systems of Sun Microsystems (NFS: Network File System), IBM, and Microsoft (CIFS: Common Internet File System). This is possible because the application software (client) and file system (server) are separate from each other, which is not the case with SAN. In an NAS the storage apparatus is directly linked to the servers, so that I/O performance is also higher.

III. 3.3 HOW do we preserve digital archive documents?

Although the tasks of digital archive management include transfer, management, opening and rendering, the issue usually boils down to the question of in what format is the digital information to be stored, and how it can be transferred to the records department. The way in which we store digital information should in theory meet a number of criteria: it must be readable, user-friendly, independent, durable, reliable, etc. Here, the choice of file format is one of the most important things, unless the archived files are to be preserved by emulation, for then the files will generally be transferred to the records department in their original format.

The file format in which data is stored depends on the application in which the file is made or consulted. An application is generally chosen for its functionality. The same thing needs to be considered when deciding the means of preserving information in digital form: which processes or operations should still be performable on the data? What actions will the system perform when that information is requested? The answer to this question will determine the format in which the data are preserved. When examining the functionality of archived files it is also important to examine whether the files are static or dynamic and to what extent they are dependent on certain applications.

Text files

When content takes precedence, text files can be stored as ASCII or SGML/XML files. If the presentation of the text is important the SGML/XML document can be given a style sheet, or, perhaps, saved as a PDF document[23]. If, on the other hand, the text file contains a number of dynamic elements that should retain their functionality (e.g. macro, index, table of contents, links, OLE object, etc.), only an application format can be used. To enable the exchange of text files with layout, the Microsoft Corporation developed Rich Text Format (RTF). The RTF specification is open. By storing a text file in RTF we can transfer the data stream and the layout data to another system. RTF is supported by many applications. Theoreticians who describe form as one of the important characteristics argue the emulation strategy because the original form can be lost when an application format is migrated. In RTF the layout is generally retained, though shifts can occur.

Audio-visual files

With audio-visual files a tool should be provided to ensure the correct decompression of the data. Almost every format in which audio-visual files are stored uses a certain method of compression to control the file size and enable exchange. To open the files the application must use the corresponding method of decompression. The records department must support the file format in which digital data is preserved and the medium on which the data is stored. Converting application-linked files to a standard or open format usually results in a loss of some data and functionality (e.g. colours). Examples of standards are GIF, JPEG, TIFF, PNG, SVG, CGM (pictures), WAVE, MP3, MP4, VoiceXML, (sound), MPEG (movies).

---

[23] The choice of PDF should certainly not rest on the alleged "inalterability" of the information. There are plenty of applications available that make it possible to alter the content of a PDF file or reprocess it in a text editor (e.g. Photoshop, Ghostscript, PDF2RFT, PDF2TXT, BCLDrake, etc.)

Databases

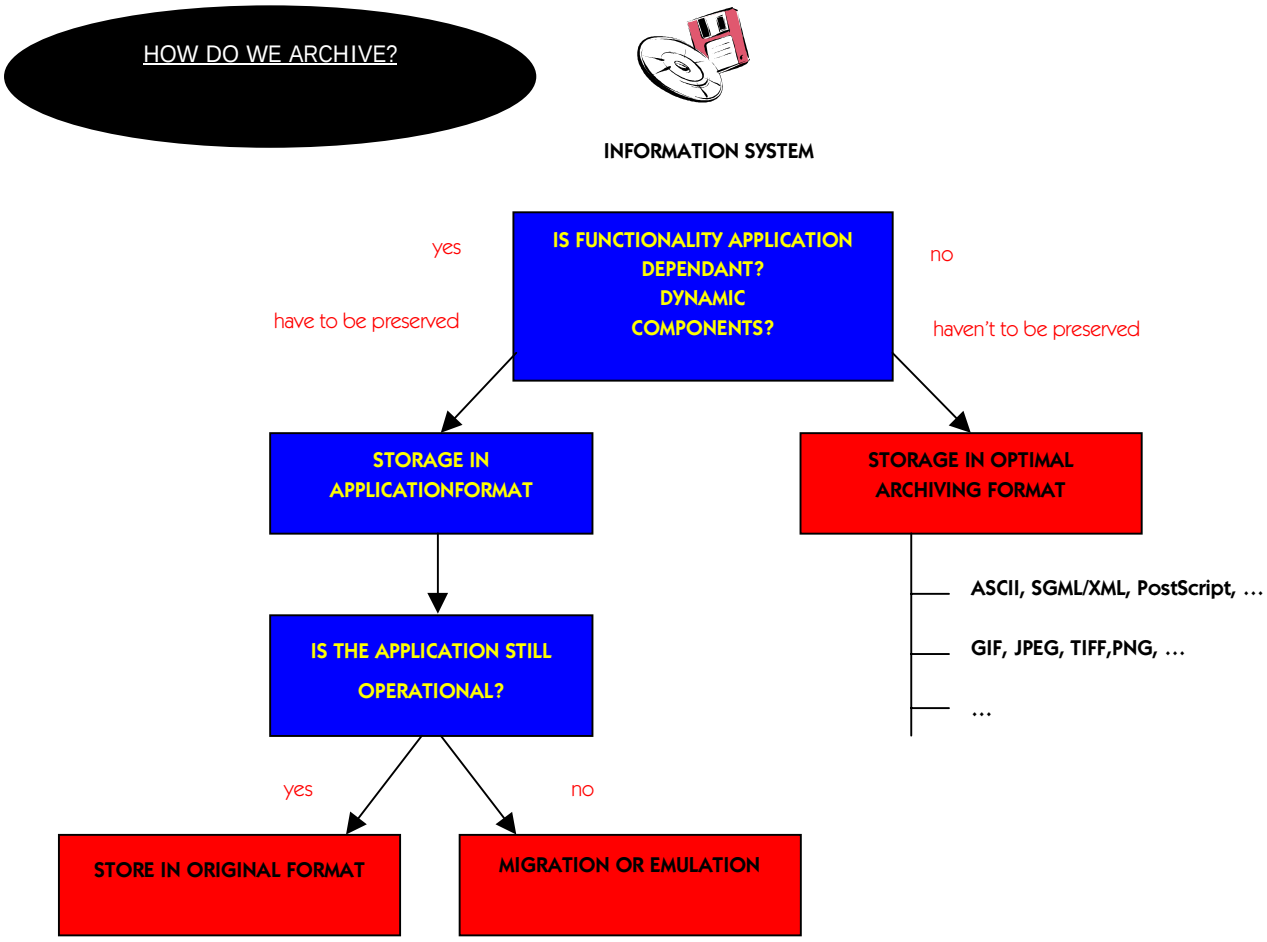
Hierarchical databases can be preserved relatively easily as flat files (ASCII, XML) with no loss of information. We just have to ensure that the structure (root-parent-child) is preserved too. XML is highly suited to this. Preserving relational and object-oriented databases is more complex. Storing relational databases in flat tables results in a loss of relations between the tables, indices and keys. With object-oriented databases one object is identifiable as one XML element. The history of a data object can be one sub-element, consisting of several elements in its turn. Both database systems can write their data to disk as XML files and manage these, but it requires a software layer above the XML files to transfer and process the data. This software layer can contain the keys, relations and indices for relational databases, and the interactions for object-oriented databases. Whether or not the software layer is preserved depends on the functionality of the archived data. It is preferable that queries be made in SQL or another standard query language.

The choice of file format is ultimately defined by the same two factors, which are central to the analysis of the information system: what functionality does the archived data have, and do the files still contain dynamic elements? In the case of static files whose functionality is supported by several applications it is best to chose an official or de facto standard. The majority of future management activities will then be migrations. In the *custodial* tradition it is recommendable that the records department draw up a list of the standard storage formats. This should contain the formats they support and the formats in which files should be stored on transfer to the records department.

Under the increasing effect of the Internet and other network applications the structure of modern information systems is becoming more open and separate. For example, the data is stored in the information system in SQL databases or XML files and can be consulted from various applications. This is closely connected to a clear partitioning of the information system layers, which simplifies long-term preservation. The way we currently migrate digital archive documents to a preservation format and the architecture used in this type of open information system are strikingly similar. Whenever functionality is confined to opening and consulting the documents, these digital data can be included in an archiving system with relative ease.

On the other hand, if the files contain imbedded programs, macros or other active components, or if they are heavily platform-dependent, or are required to retain their specific, original functionality, in most cases they are best preserved in their original environment, unless they can be easily migrated to a more suitable format for archiving, but this will prove the exception rather than the rule. In many cases the choice will be restricted to original (*native*) format or a more suitable archiving format that does not adopt all of the elements. The appropriate launch or viewer should also be preserved with the archived files (*logical encapsulation*[24]). After this the viewer should be emulated because it will only function within a specific operating system. It goes without saying that interim solutions can always be found, or emulation can switch to migration, and vice-versa.

[24] Encapsulation can also mean that all the structural information is stored in one file with the data stream: physical encapsulation.



The question of how digital archive documents are preserved also relates to the way in which they are entrusted to the custody of the archival service (*custodial*). Several scenarios are possible and each is really only a matter of practical organisation. The department can preserve them manually by putting them on a portable medium and handing them to the archivist. The records department decides on which media the digital archives will be put. In view of their digital character, archive documents can also be mailed or copied to another location. In this case there is no problem in deciding a portable medium, but a number of other aspects do arise, such as certainty over the sender's identity and the integrity of the data.

III. 3.4 WHEN do we transfer the digital archive?

The point at which we would normally transfer paper archive documents to the records department depends on the administrative storage term. When the administrative storage term has passed paper documents can be destroyed, or deposited at the archival service. In a digital context the time of transfer won't simply coincide with the end of the administrative storage term. First and foremost it is the body or administrative department taking custody of the digital archives that will determine the time of transfer. The time of transfer will be later in the *non-custodial* model than in the *custodial*

model. If the creating agency has custody of its own digital archives a file transfer will be considered once the files are no longer compatible with new versions or applications, or when there are problems relating to file size or system performance. There will probably be a selection process prior to this point. When a new system or version is introduced a choice will need to be made between migration to a software environment supported by the records department, and migration to the new application or version. In certain cases the files may be earmarked for destruction. If, however, the records department takes custody of the files once the administrative storage term passes, the transfer will take place much earlier.

The timing also depends on the type of digital archive document and the way in which obsolete data are preserved. If old information is always being overwritten because the file is processed or altered on a continual basis, and it is important that the old information be preserved, then the records will probably be archived with a certain frequency. In this way several versions (snapshots) can be preserved. This will be the case, for example, with active databases in which information is being permanently changed, supplemented or erased. If not, there is a danger that no snapshots are available. Storage should be planned for the archiving of several versions of the same file. If the content of a file is fixed and is not going to change, then in most cases only the original or final version will need to be preserved.

The increasing spread of object oriented computer applications makes it possible to manage historical information more easily in the information systems themselves. The history of an object is, as it were, a separate object within the object itself, so that it is no longer necessary to preserve several versions of the same digital information in order to trace its historical development. We can construct a current state of affairs for any particular point in an object's past using data from that object's history. The result being that it is much easier to manage historical information in the original information system (*non-custodial concept*). This type of application could involve, for example, a digital map of the city, which is in a perpetual state of change. Instead of taking snapshots of the entire map at a given frequency the data is kept on an object-by-object basis. By keying in a date the user can compile a map of all the objects present in the city at that time.

III.3.5 APPLICATION: What? How? When? Who?

All of these transfer issues came up for consideration when the Antwerp City archival service was archiving the BEAM application. The Office of Records uses BEAM to manage the register for the Public Affairs business unit. The actual data of the population register are stored in a hierarchical IMS database on mainframe. In December 1999 a new application was introduced based on a relational DB2 database. The data from the IMS database were transferred to the DB2 database. Handwritten data on members of the public prior to 1983 were not transferred. This data relates to around 83,000 people. In the future the functionality of the preserved data will be confined to retrieval and consultation. There is no need therefore to preserve the original application (top layer). There are no more data to enter or process, there are no more documents to generate, there are no more operations to perform, and a number of other tools can be used for consultation (viewer, editors, browsers, XQuery tool, etc.). The data stream is not dependent on the structural layer (middle layer) or other information systems. Obviously, the file was transferred to a platform that the archival service supports. Moreover, the computer department stopped using and supporting mainframe IMS database.

In the end, we opted for an XML file format, which is a platform and vendor-independent standard format that preserves the semantics and file data together. The files were migrated to XML in various stages. In an initial file transfer the characters were converted from EBCDIC code (mainframe) to the ASCII character set (server/PC). This gave us a flat file (ASCII) to which the beginning and end tags of the elements were added. The ASCII file was converted to XML in this way. Due to the large file size (about 350 megabytes) it was necessary to split the mainframe file into smaller files. The names were alphabetically sorted and all names starting with the same letter were placed in the same file[25].

[25] For more details (including DTD used and issue of diacritical characters) visit the DAVID website (cases → data from the population register).

IV. CONCLUSION: ARCHIVING REVISITED

When dealing with digital archive management it is best to start from the information system itself. No typology, whether it rests on the function of the documents generated or the computer or application file type, can serve as the basis for a preservation policy. The file type (text, spreadsheet and database), type of information system (mainframe, server and/or client application), form, functionality and file format of a digital archive document is variable. Though the function of a digital document may be well rooted in the work process, it cannot be used as the basis for developing an archiving strategy.

When it comes to digital file preservation, migration and emulation are the options being explored and indeed applied today. Whichever of these strategies is most suited will depend on the characteristics of the information system and the future use of the archived data. The best way to elucidate these factors is to divide the information system into layers (data stream, structural information, application). In the first place, migration is about preserving the data stream and converting the structural information if necessary. Whenever there is a need to preserve the structural information and/or tools, the emulation strategy comes into play. A crucial factor in the choice between migration and emulation is the future processing demands that will be made of the archived data, and the matter of whether their functionality can be supported by an independent platform. Since this is a matter of study for every separate information system it is particularly important that every case be documented. At the very least the data recorded on each information system should make it possible to ascertain the archive value. A more detailed technical description can then be made of the information systems with archive value (see annexe 1). The necessary metadata can be taken from this inventory of digital information systems at a later date.

Migration and emulation each have their advantages and disadvantages. In the international literature on the subject some authors set out their stalls for or against a given strategy. For a records department with the task of preserving a huge variety of information systems this discussion has no relevance. Either strategy is suited to the preservation of digital information from a given information system. It is safe to say that migration can be applied whenever the content of a digital file is fixed and when its functionality is supported by several applications. With migration, the main thing is to archive the data in the best possible format for preservation. In cases where the digital information depends on a specific application and we need to preserve the dynamic properties too, emulation is the better option.

Digital preservation does not merely imply the durable capture of the bit streams contained in an information system's data. These data are created in an information system. The details on this information system and the data itself need to be preserved, for these cannot (in all likelihood) be derived from the digital archive documents in themselves. In this sense it is not just archive documents that are preserved, but information too.

Since digital preservation is not document-based, we need to look again at some other aspects of archiving work, such as description, preservation of condition, selection and rendering. With our present terminology and description techniques, describing digital files is only ever a straightforward matter in cases where the digital information has a paper equivalent. We can describe a word processing file containing a letter or annual report in the same way that we would describe a letter or annual report on paper. All we need change in the description is its material form, and then add a few

technical details. Descriptions are more difficult when there is no paper equivalent, when preserving an entire information system that produces several types of documents, or when no documents are produced and the system is only a manager of information. One solution would be to record the file type (text, spreadsheet, database, graphic, music, video file) and add to this a description of content. The work process will be an important descriptive element of any information system, emphasising yet again the importance of this field in an inventory of digital information systems.

When selecting digital files it is necessary to consider the technical feasibility of destroying (that part of) the digital information, without bringing the information system to a standstill or rendering illegible the part to be preserved. Paper documents are physically separate from each other whereas digital information is often linked. Deciding what to destroy raises the same questions as deciding what to preserve. In a digital context this implies selection at the system level. It will rarely be possible to use different storage times for archive documents from the same digital information system. Consequently, the selection lists will need to be amended.

Finally, we should bear in mind that it takes hardware and software to render a digitally preserved file. This equipment will need to be provided in the reading room. In addition to the archived files, appropriate search and view tools will be required. If the archive-creators has custody of its digital information, this information must be open and available to researchers.

A strict policy and structured approach is particularly important in digital archive management. This report gives three steppingstones to a digital preservation policy. In the inventory of digital information systems we have an overview of the digital information systems present, a source for the metadata we require, and a method of ascertaining archive value. Every information system breaks down into three layers. The decision model tells us which information layers are preserved, who manages the digital archive, when the transfer is made, and how the archiving is done.

ANNEX 1: SPECIMEN INFORMATION SYSTEM FILE

NAME Name of application (+ meaning of abbreviation)	
PURPOSE Describe the purpose of the work process in which the system is used.	
FUNCTIONS How does the information system function? What are the main functions? What dynamic components are contained in the computer files?	
DEPENDENCIES: DATA Is the system linked to one or more other systems? Which? **DEPENDENCIES: TOOLS** Is the system dependent on certain tools? What is the function of the tools? Are there other tools that support this functionality? Which ones?	
PLATFORM What hardware and operating system does the information system use?	

HISTORY Previous information systems? Date of - commissioning? - closure/transfer? Versions? Migrations of the digital information? Successive information systems?	
DEPARTMENTS Which department(s) manage/use the information system?	
INFORMATION Which digital information does the system generate or manage? Public accessibility to the information?	
FILES Which storage formats are used? Volume?	
FUNCTIONALITY OF THE ARCHIVED INFORMATION Which operations or processes should be possible on consultation?	
ARCHIVE VALUE What is the archive value of the digital information?	

ANNEX 2: DIGITAL ARCHIVE CHECKLIST

There is a whole decision-making process to go through before preserving digital archive documents. A checklist of the main points for consideration can be a valuable aid when preparing digital archives or setting up a digital preservation system. Below, we summarise a number of elements to bear in mind when:

➢ Developing and using digital information systems

- ✓ Is there supervision to ensure that standards are implemented as much as possible in the development of digital information systems?
- ✓ Do the information systems have an open structure? Can data and structural information be separated from each other?
- ✓ Is the archival service regularly informed of the development of new information systems or the alteration of existing information systems (e.g. duty to report)? Is the archival service consulted?
- ✓ Is the management of the digital information systems documented (metadata)? Is an overview of the digital information systems kept? Does the administration know which metadata keep up-to-date?
- ✓ Does the department have guidelines on working with digital (archive) documents (storage, destruction, alteration, etc.)? Is there a clear understanding of what is archive material and what is not?

➢ Digital preservation by the agency creating the archives:

- ✓ Can the digital information system manage, open and render the archived information? Are special functionalities or modules provided for this? Is it possible to select?
- ✓ Who has custody of the preserved digital archive documents?
- ✓ Can archive users consult the digital archive documents?
- ✓ Have steps been taken to guarantee authenticity and integrity?
- ✓ Is there a procedure in place in case the information system is altered?
- ✓ Is there a procedure for making and managing security backups?
- ✓ Has a time been agreed for depositing (based on file size, upgrades, loss of support, etc.)?
- ✓ Are snapshots taken whenever the old data are systematically overwritten? What is the periodicity of snapshots? Are the snapshots transferred to the archival service?

➢ Digital preservation by the records department?

- ✓ Is there an overview of the quality requirements to be met by the digital archive documents deposited (file format, medium, file system, file size, character code, etc.)?
- ✓ Does the records department have the hardware and software needed?
- ✓ Are the digital archive documents stored in a file format and on a medium acceptable to the records department? Is an emulation program provided? Does the file size constitute a problem for the server, PC or network?
- ✓ Have the necessary queries, style sheets and log files been deposited? Is there a need to provide access?
- ✓ Is there a deposit procedure?
- ✓ Are the requisite metadata present?
- ✓ Have the migrations been documented?
- ✓ In the case of non-current information systems, does the archival service have a users' manual?
- ✓ Can compressed files be properly decompressed? Is the software required for this to hand?
- ✓ Does the archival service or institute possess the necessary software licenses?
- ✓ Can the preserved digital information be included in the archive system?
- ✓ Have steps been taken to guarantee authenticity and integrity?
- ✓ Is a procedure in place for making and managing security backups? Are the digital files regularly transferred to another medium (refreshed)?