Digital containers for shipment into the future

Filip Boudrez
Expertisecentrum DAVID vzw
Antwerpen, 2005

0. TABLE OF CONTENTS

1. INTRODUCTION

Now that electronic records management and electronic record-keeping are being implemented more and more, the day that electronic records will be transferred to archival services or institutions is approaching. Ideally, these electronic records should be stored in specially designed digital repositories.

At present, various research and pilot projects are being conducted with digital repositories as scope[1]. Because of the special nature of electronic records – which are so much more than merely digital objects[2] – special requirements apply for the organisation, the functionalities and the workflow of a digital repository. An important point, that one definitely may not forget, is the digital durability of the repository itself. Each information system, and thus also a digital repository, is subject to technological obsolescence. A sustainable storage method must be able to survive the obsolescence of particular hard-ware or soft-ware components used for the digital repository. In designing and implementing a digital repository, the archivist must take this into account.

The complexity of this issue is directly related to the way in which the content of the digital repository is organised and composed. The archivist can make allowances for this problem if he anticipates into this during the selection of a storage method for the electronic records and their metadata. A well-chosen storage method can keep the long-term management of the electronic records

[1] Examples of initiatives by national archives include: National Archives of Australia (AtoR Project), National Archives and Records Administration, Schweizerisches Bundesarchiv (Arelda Project), etc. A project closer to home is the e-Repository of the 'Gemeentearchief Rotterdam' and the Archiefschool.

[2] See for this: K. Thibodeau, *Boundaries and transformations: an object-oriented strategy for the preservation of electronic records*, in: *Proceedings of the DLM Forum on electronic records*, Brussels, 1996, p. 161-167; F. Boudrez, *A. Introduction, The electronic record*, in: F. Boudrez, H. Dekeyser and J. Dumortier, *Digital archiving: the new challenge? Legal and archival issues,* Mont Saint-Guibert, 2005.

uncomplicated, both intellectually and technologically, protect the electronic records better from calamities and minimizes risks to an absolute minimum.

In this article the storage method is expounded that eDAVID developped and that will be applied by the City of Antwerp. This storage method expands on the digital preservation strategy for electronic records recommended by the DAVID-project. Thus this article starts with a description of this preservation strategy. Without a clear view of the preservation strategy to be followed, it is virtually impossible to organise the content of the digital repository in a structured manner. In the second part of this article attention is mainly given, not only to a storage method that is suitable for this digital preservation strategy, but also one that offers a solution for the preservation of electronic records in relation to their metadata. For this, the approach of the Open Archival Information System (OAIS)[3] is used. Particular emphasis is placed on the composition and the management of *Archive Information Packages* (AIP's). In the third part of this contribution, a model structure for an AIP is presented. Then the XML Schemas are presented that have been designed by eDAVID for the City of Antwerp. XML will be used extensively as an archiving, metadata and encapsulation format in their digital repository. These XML Schemas serve as an implementation example and are documented in the fourth part of this contribution. Finally, this article concludes with a general comment on the implementation of the presented storage method.

2. THE DAVID PRESERVATION STRATEGY

Before the archivist can determine how to organise digital information best in the digital repository, he must have an insight into the way the electronic records will be reconstructed on the screen in future on the basis of the archived bits and bytes. The digital preservation strategy that an organisation applies will determine which information is archived and how it can be managed in the digital repository.

The DAVID project[4] evaluated existing digital preservation strategies for digital objects and examined the extent to which they are suitable for archiving electronic records in the long term. This meant, among other things, that consideration was not only given to how electronic records can remain renderable with the passage of time, but that an investigation was also made as to how the authenticity and interpretability of the electronic records can be guaranteed.

This research indicated that at present migration and emulation are potentially the most suitable digital preservation strategies[5]. Much ink has flown about the advantages and disadvantages of both strategies[6], but in essence, migration and emulation do not exclude each other. As a matter of fact, both strategies actually might be complementary: each strategy is most suitable for certain types of records. For textual documents without any functionality, migration is no doubt the simplest preservation strategy, whereas for dynamic records with certain behaviours (f.i. websites), emulation might be the most advisable option. Archives can apply both strategies for different document types. During the life cycle of a record, a migration and an emulation phase can also follow each other. Emulation has the best chance of success when records are stored in an open, documented and standardised file format. This could mean that records in a closed proprietary format first of all have to be migrated to an open file format before they can be emulated.

---

3 ISO-14721(2003): *Space data and information transfer systems. Open archival information system. Reference model.*

4 http://www.edavid.be/davidproject

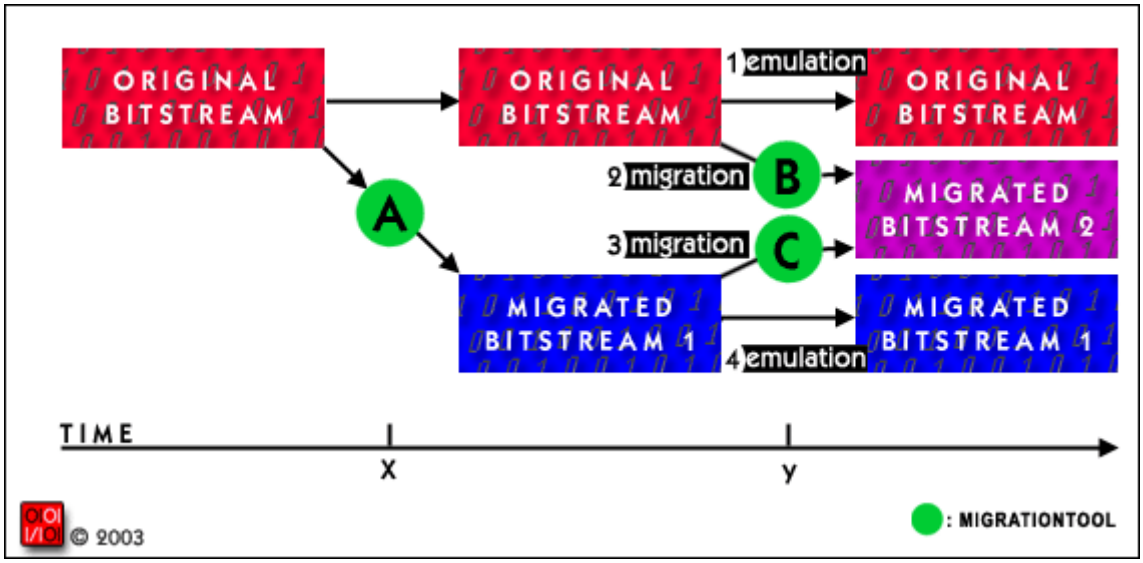
5 F. BOUDREZ, *B. Preservation strategies,* in: *F. BOUDREZ, H. DEKEYSER AND J. DUMORTIER, DIGITAL ARCHIVING: THE NEW CHALLENGE? LEGAL AND ARCHIVAL ISSUES,* MONT SAINT-GUIBERT*, 2005.*

6 See for example, J. ROTHENBERG, *Ensuring the longevity of digital information*, 1999 (http://www.clir.org/pubs/archives/ensuring.pdf); D. BEARMAN, *Reality and chimeras in the preservation of electronic records*, in: *D-Lib Magazine*, April 1999 (http://www.dlib.org/dlib/april99/bearman/04bearman.html)

To offer as many readability guarantees as possible and to keep both the migration and the emulation option open, the records are ingested in the digital repository in both their original file format and in their suitable archiving file format. Each format can be considered as a different representation of the same record. In practice, however, there will not always be precisely two digital objects archived for an electronic record. When the original format is directly the most suitable archiving format for that document type, only one digital object for the record will be included in the repository. On the other hand, it is also possible that for a certain type of digital documents, several archiving formats might be used, resulting, for example, in three digital objects being archived.

If in future the required hardware and software are unavailable for consulting the original format and the archiving format, with the DAVID preservation strategy one has at least four options for reconstructing the records on the screen:

- emulation of the original format
- migration of the original format
- migration of the suitable archiving format
- emulation of the suitable archiving format



By including both the original and the migrated bitstream in the digital repository, one anticipates also the future technological evolution. The information technology continues to evolve quickly and it is not possible to predict what will be possible in the future. After a time, one might wish to consult electronic records with a technology that was not yet available at the time of ingestion in the repository.

This preservation strategy, in addition to providing as many readability guarantees as possible, also offers some other advantages. Firstly, users can consult an electronic record in both the original bitstream and in a migrated bitstream, depending on their preference or on the software applications they have. Secondly, when the original bitstream is archived, authentication remains possible on the basis of technologies that relate to the original bitstream. Advanced digital signatures are an example of this. A condition is that all elements of the 'validation chain' and the necessary metadata must be available[7]. Thirdly, records in their original and migrated bitstream can be compared or the migration process can be reconstructed.

[7] F. BOUDREZ, *Digital signatures and electronic records,* Antwerp, 2005. (http://www.edavid.be)

3. ENCAPSULATED ELECTRONIC RECORDS

In the DAVID preservation strategy, one or more representations of the same electronic record are preserved in the digital repository. In addition to the representations, metadata about the record and its representations must also be captured and archived. The archivist has various options for the storage of the various representations and the metadata of an electronic record. The organisation of the digital repository depends largely on what the archivist selects as basic unit, and on the way the records are identified and their metadata are stored.

3.1 Encapsulation as a storage method

An extremely important point in the selection of a storage method are the identification and the relationships between the various components in the electronic record. The components of a record form a logical entity which of course may never be lost after it is transferred to the archives.

With most storage methods, the various components of an electronic record do not form a physical entity, but are stored at separate locations (in a database, a file system or a combination of both[8]) and as different digital objects. Their mutual relationship is indicated by means of links, database relations, pointers and filenames. Archiving these relationships is not self-evident in the (medium) long term. The fast evolution of information technology requires that the relationship between the digital objects be established in a clear and permanent manner. This is not an insurmountable problem, but it is an important point and can involve a challenge as time passes. In addition, the danger always exists that the relationship might be lost.

Preserving the components of an electronic record separately always involves a risk. As soon as the mutual relationships are broken and cannot be reconstructed, the record must be considered as lost. Metadata are indeed essential for the recordness and the long-term readability of the electronic record. The archivist can avoid this risk by including metadata in the computer files that contain the documents. By combining both components in one physical object, the relation between the record and its metadata is prevented from becoming lost.

The addition of metadata to digital objects is called encapsulation or embedding. Encapsulation is sometimes mentioned in one breath with migration and emulation, as a digital preservation strategy[9], but strictly speaking it does not belong in this category. Encapsulation is actually nothing more than a storage technique in which metadata is added to a digital object and/or several documents are grouped in one digital object. Encapsulation is not a method that prescribes how digital documents will be reconstructed on the screen in future or how accessibility is preserved.

The idea of embedding metadata in the digital objects is not new. Encapsulation is one of the basic principles of object-oriented programming. The importance of encapsulation for electronic record-keeping was actively promoted in 1996/'97 by David Bearman[10]. In the meantime, encapsulation has been applied, for example, in the *Persistent Object Preservation* of the NARA, in the VERS archiving

---

[8] Possibilities are:
– metadata: database / representations: database (BLOB's)
– metadata: database / representations: file system
– metadata: file system / representations: file system.

[9] For example, http://www.nla.gov.au/padi/topics/18.html; TESTBED DIGITALE BEWARING, *XML for digital preservation*, The Hague, 2002.

[10] See for example: D. BEARMAN, *Item level control and electronic recordkeeping,* in: *Archives and museum informatics,* vol. 10 (3), p. 195-245; MARK D. GIGUERE, *Metadata-Enhanced Electronic Records*, Philadelphia, 1997.

strategy of the Public Record Office of Victoria and in the AtoR Project of the National Archives of Australia[11].

The enrichment of records with metadata is not an absolute prerequisite for permanent electronic record-keeping, but it is well worth considering since it provides important advantages:

- the metadata form a part of the archived digital object and are not stored at an external location. The metadata are inextricably connected with the record. One does not have to worry about links or pointers between digital objects and their metadata. Encapsulation also facilitates management in the (medium) long term.
- all components of an electronic record can easily be transferred and migrated together.
- the electronic records are self-descriptive and autonomous: they identify and document themselves
- the embedded metadata can be extracted at any time and stored centrally
- the digital objects in the digital repository have record status without needing external information. Electronic records rather than digital objects form the basic units of the digital repository
- the consequences of disasters might be less serious (risk assessment):
  - the digital repository still contains records
  - metadata can be extracted from the records.

The archivist can apply encapsulation as a storage method in various ways. A first method is to include metadata in the computer files that contain one certain representation of the record (for example, by filling in the metadata tags in the header of a TIFF file). However, this encapsulation method does not help the archivist much further. For records with several representations, the same metadata must be stored multiple times and the various representations must be related to each other in one way or another. In the technical area there are also several disadvantages to this encapsulation method. The encapsulation of metadata does not result in many problems with plain text files, but for binary file formats, this is much less obvious. Most binary file formats do provide several standard fields for the inclusion of metadata, but they do not satisfy all archival needs. In the metadata fields provided, there is usually too little space and an expansion of the fields could cause interchangeability and readability problems. The addition of metadata to binary files also requires a separate module or software tool for each format, because usually such a functionality is not supported by current computer programs.

To avoid these disadvantages it is better for the archivist to use a different encapsulation method, more in particular, the encapsulation of the various representations of the same record in one computer file. Then the various representations no longer have to be linked to each other because their relationship is indicated by the fact that they are part of the same physical object. This encapsulation method also offers the advantage that the archival descriptive metadata only has to be included once. The metadata no longer have to be embedded in the header fields of the binary file formats, but they can simply be included in the computer file. In this scenario, the archivist combines the two encapsulation options: metadata are added and various digital objects are included in the same computer file. This encapsulation method results in what are called container files[12]. One container file contains all components of one electronic record and forms the basic unit of the digital repository.

When encapsulation is used as a storage method, electronic records must be packed in container files before ingestion in the digital archive, and they can be consulted only after they are unpacked. Thus the formation of container files requires at least two additional steps in the workflow of the digital repository.

---

[11] K. THIBODEAU, R. MOORE and C. BARU, *Persistent Object Preservation: Advanced computing infrastructure for digital preservation*, in: *Proceedings of the DLM Forum on electronic records,* Brussels, 2000, p. 113-118; http://www.prov.vic.gov.au/vers/standard/advice_11; http://www.naa.gov.au/recordkeeping/er/guidelines/10-preservation.html#pres6.

[12] The physical or logical structures in which the bitstreams of a record and the metadata are brought together, are usually designated by the terms “containers” or “wrappers”.

3.2 Records as Archive Information Packages

When developing a storage method for electronic records in a digital repository, the OAIS standard is a suitable framework. In particular, the information model of this generally accepted ISO standard is a valuable reference. In the OAIS model, information packages compose the basic unit of the digital repository.

OAIS describes the functions, the processes and the information flow of a digital archive. It focuses on the information packages that are included, managed and consulted in the digital archive[13]. In the OAIS model each *Information Package* is a conceptual container that consists of two types of information: *Content Information* and *Preservation Description Information*. The *Content Information*, in addition to the bitstream of the actual digital information source (the *Data Object*), also contains its *Representation Information*. With the help of that *Representation Information, a Data Object* is transformed into an understandable *Information Object*. Thus the information packages contain various metadata in addition to the actual bitstream of the archived digital object:

- the *Representation Information*: all information that is needed to translate the digital object to an interpretable concept (for example, source code or the compiled installation files for a viewer).
- the *Preservation Description Information*:
  - unique identifiers ("reference information")
  - information about the context ("context information")
  - information about the origin ("provenance information")
  - information for the integrity or the validation of the contents ("fixity information")[14]

The archivist does have to take into account that in the OAIS model an information package does not necessarily coincide with one record. OAIS is designed for the long-term archiving of digital information in general. One AIP can correspond with one record, but also with one component of a record or with several records. In the OAIS model, various kinds of AIP's can be combined. This enables the archivist to select different aggregation levels. Together with the relevant metadata, an AIP can have as content:

- one component of a record: for example, an image inserted in a text processing document
- all components of one representation of a record: for example, a text processing document (including any images and the text)
- all representations of one record: for example, a text processing document in its original and archiving formats
- all parts of a collection or record group: for example, all records that are part of a file or subject folder[15].

Since the archiving of electronic records is the point of departure, it goes without saying that the archivist selects all representations of one record as the aggregation level for an AIP. In addition to the bitstreams of the digital objects, metadata is also included in the AIP: technical metadata about each representation and archival metadata that relates to the aggregation level, in this case, the record.

---

[13] In addition to the AIP's, the OAIS model also distinguishes two other Information Packages: the Submission Information Packages (SIP's) and the Dissemination Information Packages (DIP's). The SIP's are transferred by the creator to the archivist who transforms them into AIP's. On the basis of the AIP's, DIP's are distributed among the users.

[14] ISO-14721 (2003): *Space data and information transfer systems -- Open archival information system -- Reference model.*

[15] The *Persistent Object Preservation approach* of the NARA provides this possibility. Then the archival bond between the items of a file is registered in a physical manner. This approach has the disadvantage that the containers quickly become very large (> 150 MB) and are no longer as easy to process. An alternative to physically recording the mutual relationship is a reference in the AIP for the electronic records to the dossiers of which they are a part.

3.3 AIP's in containerfiles

In the OAIS model, information packages are considered as conceptual containers: one information package can be spread over several digital objects and this whole forms a logical unit. If the archivist wants to make full use of the encapsulation advantages, then it is advisable to store the AIP's as one digital object or one physical container file. Then one container file functions as a complete AIP and also contains, in addition to the metadata, all representations of an electronic record.

These AIP's or metadata enriched records form the basic unit of the digital repository. The option of combining all representations and the essential metadata of one record in one AIP and one computer file, corresponds best with the concept that an electronic record can have various representations, and it offers the advantage that the shared archival descriptive metadata only have to be stated once.

3.4 Metadata enriched records

The purpose of enriching electronic records with metadata is to make them as self-descriptive and as autonomous as possible. Metadata are indispensable for the management and preservation of electronic records. They have various functions: intellectual control, administrative management, archival description, demonstrating authenticity, long-term readability, preservation, supporting retrieval, etc.

A general metadata standard for electronic records is not yet available. The ISAD(G) standard for the archival description provides some few fields in which digital characterictics and properties of electronic records can be registered, but the ISAD(G) metadataset is mainly intended for description and making them accessible. Thus ISAD(G) is too limited to be used as a general metadata set for electronic records since ISAD(G) does not cover all metadata functions for electronic records. ISAD(G) can of course be used for the description of electronic records, but for the other metadata functions, additional metadata are needed. Electronic records, as digital objects and as records, need additional metadata for their long-term management.

Because of the nature of a record, an electronic record needs metadata related to records management: metadata about the context in which the record was created and received, and metadata about the procedures by which it is managed. The essential metadata must at least indicate the identity and the unique nature of the record so it can be distinguished from other records. These metadata do not have an identifying function alone, but are also important for prooving the authenticity of the record, so it can serve as evidence. These metadata also relate the records to work processes and other documents so they can be interpreted by users. In common file systems or records-management applications they are usually stored separate from the electronic records, although they are essential for the status and function of the records[16]. As an additional refinement of the records management standard (ISO-15489), a standard for metadata records management is currently being worked out (ISO-23081). Pending the establishment of this standard, existing records management metadata sets (Australia, UK, etc.) can serve as a source of inspiration.

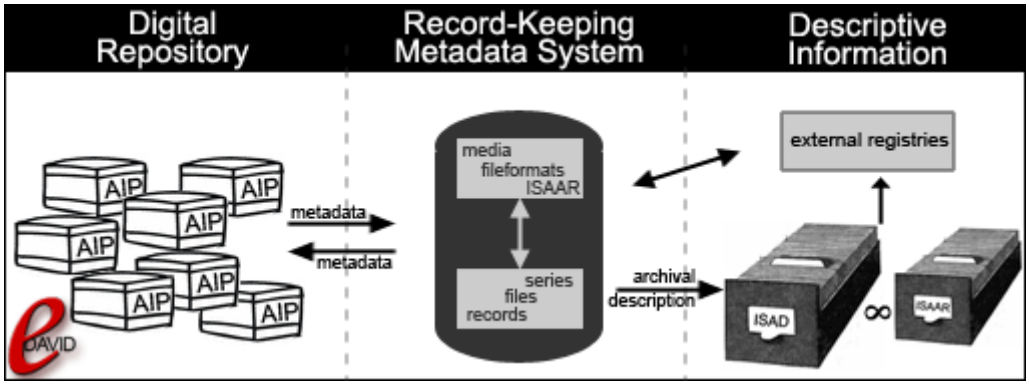
In addition to the extra metadata from a purely archival perspective, several other technical metadata about the digital representation(s) of the electronic records also qualify for encapsulation. These technical metadata provide documentation about the preserved bitstreams from which the records must be reconstructed on the screen. Without any technical documentation it will be difficult, if not

[16] The importance of metadata for the '*recordness*' of electronic records, and of the encapsulation of metadata, was already emphasised in 1996 by David Bearman. The project 'Functional Requirements for Recordkeeping' views a record as an object with embedded metadata and compares a record with a document stored in an envelop that is labeled with the necessary metadata. (D. Bearman, *Virtual archives*, at: *ICA Meeting*, Beijing, September 1996 (http://www.archimuse.com/papers/nhprc/prog6.html).

impossible, to consult electronic records after time has passed. The technical metadata support the reconstruction of the archived bits and bytes into a human readable document on the screen. Which technical data are essential depends on the type of record and the preservation strategy that is being used.

With regard to technical metadata it is interesting to look beyond the limits of the archival science at initiatives such as OAIS and Premis. The OAIS reference model prescribes which types of metadata are needed for the archiving of digital information, but the OAIS standard does not indicate exactly which metadata fields are needed. The archivist can find this included in the metadata set for the preservation of digital objects developped by the Premis Working Group. Expanding on the OAIS information model, the Premis Working Group has identified the metadata that are needed for the archiving of digital objects[17].

A crucial matter for the application of encapsulation as a storage method is determining which metadata will be included in the container files. Technically it is possible to embed all relevant metadata, but this is not always efficient or expedient. First, one must take into account the incremental nature of the metadata process. So the embedded metadata will not have to be up-dated frequently, it is better to include only the essential metadata. Second, encapsulation of all metadata leads to overkill of the container files and to a huge redundancy in the digital repository. It makes little sense, for example, to include in each container the specifications or viewers for certain file formats (*Representation Information*) or documentation about the creator (ISAAR authority record). Such shared metadata only have to be preserved one time. Also, the encapsulation of metadata does not mean that metadata will no longer be stored centrally or externally, to the contrary.



The central record-keeping metadata system (RKMS) can contain various metadata: for example, authority records with descriptions of creators (ISAAR), documentation about file formats (represenation information), management information about the storage media, archival descriptive metadata at a higher level than records (for example, dossiers, series, record groups, etc.). Some of this metadata might be extracted from or linked to external registries. Even the metadata embedded in the AIP's can be repeated in the central record-keeping metadata system (for example, to make faster search actions possible). Because of this, synchronisation is sometimes needed between the AIP's and the central metadata system, but this also serves as an additional security measure. Ideally, on the basis of the metadata in the general metadata system, ISAD(G) compliant archival descriptions can also be (automatically) compiled.

As a consequence it is best for the archivist to be selective in choosing the metadata that will be embedded. When selecting the essential metadata to be embedded, the archivist must seek a balance

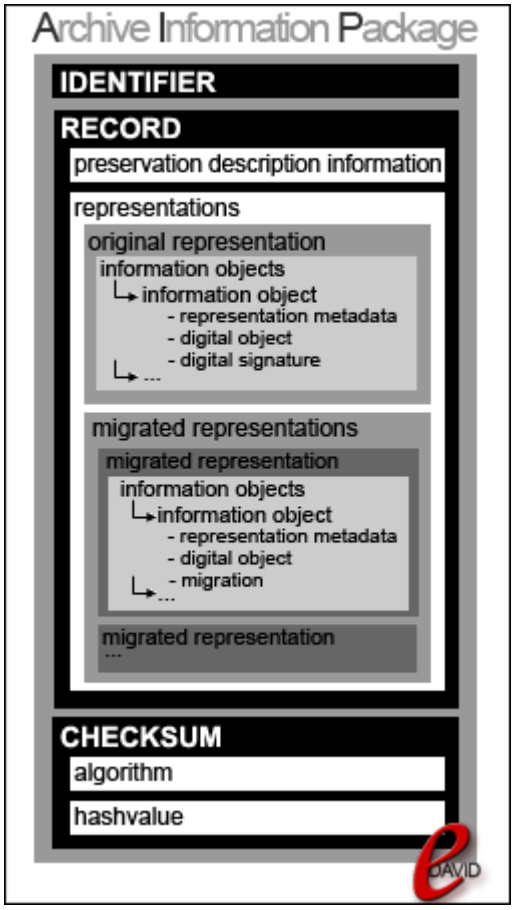
[17] PREMIS, *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, May 2005 (http://www.oclc.org/research/projects/pmwg/premis-final.pdf).

between autonomy on the one hand and keeping the container files manageable on the other hand. Criteria for this selection are:
- metadata at record level
- metadata that are essential for the recordness and the interpretation by humans and machines.

4 A MODEL AIP

An *Archive Information Package* that, in addition to all representations, also contains the essential metadata of a record, can be modelled in various ways. Below, a hierarchical model structure is presented and an explanation is given as to how a complete AIP can be stored in one computer file.

4.1 The AIP-structure



The main structure of an AIP consists of three parts:
- the identifier for the AIP
- all representations and the essential metadata of the record
- the checksum.

The identifier and the checksum serve mainly for the management of the AIP's. The identifier contains the unique ID of the computer file with the AIP as content and is the reference to the AIP. Preferably, this should be a permanent ID so it can serve as an identifier for the AIP on a long-term basis. The checksum functions as 'fixity information' and can also be used as (part of) the AIP identifier. With a checksum, the validity of the AIP's can be thoroughly checked afterwards by comparing the embedded and the recalculated hash values with each other. This check can be carried out completely automatically and randomly. If the embedded hash value is not equal to the recalculated hash value, an alarm function can be activated (for example, to retrieve a backup). For the checksum, not only the hash value is preserved, but also the applied hashing algorithm.

The second part in this AIP structure contains all components of the electronic record and is split further into several subelements. The archival descriptive metadata and the records management metadata are included in the subelement 'preservation description information'. These metadata relate to every representation of the electronic record and therefore only have to be stored once. The second subelement ('representations') contains all representations and the technical metadata of the electronic record. The structure provides space for one or more archiving file formats besides the original representation of the record. A record can have more than one suitable archiving formats or, in future, new migrations can be needed. Each representation may consist of one or more computerfiles ('information objects'), as there might be a one-to-one or a one-to-may relationship between a record and computerfiles.

For the original representation, a place is also provided for an optional digital signature and all associated metadata. This element is provided for the archiving of digitally signed records and will in principle only appear as a subelement of the original document. In this element, in addition to the digital signature and its metadata, all essential elements of the validation chain can also be included[18]. In addition to the technical metadata and the actual digital object, there is also space for documenting the migration operation ('migration').

4.2 XML as container file format

The containers with the AIP's can be stored as one digital object in various ways. As a file format for the container files, XML is preferable to compressed 'archiving' formats such as tar, zip, gzip, jar and gz.[19] These formats should be avoided because of the additional reconstruction link that results from the use of compression. Digital files that consist exclusively of text characters can be included directly in XML. Binary files must first be converted to text characters using Base64. The packing and unpacking via Base64 can also be viewed as an additional reconstruction link, but Base64 is very well documented, widely distributed on various platforms and very simple. Converting binary files by means of Base64 has the opposite effect on the file size as does compression. The file size increases by about one third.

A second reason for not using compressed 'archiving' formats, lies in the limited possibilities for including metadata in these formats[20]. Since XML can be expanded, one can provide metadata fields in the XML containers depending on one's own needs. Regardless of the format, the metadata in an XML container can be processed in a uniform way with one software component. The inclusion of metadata in the XML containers offers the advantage that it is no longer necessary to have a persistent link between the digital objects and the metadata which are stored separately.

The embedding of metadata in the XML containers not only offers the advantage that the metadata are inextricably connected with the record, but it also ensures that the metadata are recorded in a structured and static manner so they can be easily used again afterwards. By keeping the metadata in XML, one also ensures that these data are recorded digitally in a permanent manner and can easily be processed by computers. Finally, the flexible nature of XML makes it possible for the elements in the XML container files to be expanded or transformed at any time. This is another advantage with a view to the incremental description of records or subsequent migrations.

The selection of XML as the file format for the AIP's is responsible not only for technical reasons or for the sake of long-term consultability, but also because of the need for documenting the AIP's. Information and knowledge about the composition and the content of the AIP's must be transferred through time. It must be possible for future generations of archivists to understand how the representations and the metadata of an electronic record are packed in a container. The semantics of the AIP parts and their mutual relations must therefore be documented in a clear way. With XML, the composition of the AIP's can be communicated in an orderly and well-structured manner, without it being necessary to consult external information. In OAIS terminology, the XML Schema for the AIP will function as *Packaging Information* (the information that connects the parts of an information package with each other). The XML Schema for the AIP identifies and relates the components of the *Content Information* and the *Preservation Description Information*. This *Packaging Information* is largely embedded in the AIP's in the form of semantic and nested XML elements. The AIP's are thus self-descriptive. Using XML for the AIP's provides the same advantages as the embedding of metadata in

[18] F. BOUDREZ, *Digital signatures and electronic records*, Antwerp, 2005 (http://www.edavid.be).

[19] F. BOUDREZ, *<XML/> en digitaal archiveren*, Antwerp, 2002. (http://www.edavid.be)

[20] A solution for this limitation is the customisation of the metadata files in the compression formats. An example of such an initiative is research into the customisation of JAR files so metadata can be included (W.E. UNDERWOOD, *A java JAR implementation of an archival information package*, Consultative committee on space data systems, XML Workshop, NASA Goddard, 20 August 2001).

the records. A condition for this is that the document model on which the AIP is based, is constructed logically and in a structured manner, and that XML semantic element names are used.

When using XML as the AIP format, rather than just using 'normal' XML, the archivist might consider using XML in conjunction with the Resource Description Framework (RDF) model. Just like XML, RDF is a W3C Recommendation[21] and is designed mainly for the semantic web. RDF increases the machine readability and the interoperability of the XML documents. On the other hand, the semantic web is not the main concern of archival institutions and by the application of RDF syntax, the XML structure becomes more complex. RDF requires a different method of information modelling. This is at the expense of an easy readability by humans because the interpretation of the structure and the distillation of the semantics is more difficult.

5. XML SCHEMAS

By using XML as the file format for the container files, each organisation can work out a custom-made container model for the AIP's depending on its own needs and approach.

For the implementation of the above-described storage method using XML container files, eDAVID developped various XML Schemas for the City of Antwerp. These XML Schemas define the formal model for the XML documents. There is an XML Schema for:

- the XML container file or the AIP
- a general record-keeping metadataset for the management of electronic records (work in progress)
- the archival descriptive metadata in conformity with ISAD(G)
- the document types for which XML is used as the archiving format: e-mails, calendars and databases.

This strategic choice of XML results in a combined application of XML. First, XML is used as a language in which all parts of an AIP are packed as electronic records. Here XML is used as an encapsulation format. Second, XML is also used as a suitable archiving format for several document types. Third, XML is also used as the metadata format for the essential metadata. These metadata are stored directly in XML.

When developing the XML Schemas, eDAVID did not decide on the use of RDF, but on 'plain' XML. In addition to the above-mentioned disadvantages, it was not considered to be desirable to embed a time-bound technology such as RDF in the basic units of the digital repository. To preserve surveyability when combining various XML Schemas, different namespaces were applied in the XML Schemas. All XML elements defined in the AIP XML schema have the prefix 'aip'. The elements in the general record-keeping metadataset have the prefix 'rkms'/ The elements from the ISAD(G) description standard have 'isad' as prefix, whereas in the XML Schemas for the document types in which XML serves as the archiving format, the prefix refers to the document type (e.g. e-mail, calendar and database).

The design Schemas were the subject of a 'request for comments' that was sent to various (inter) national archival institutions and colleagues on 30 May 2005. Their reactions were collected by 30 June 2005 so their comments could be processed in July and August 2005.

[21] W3C, *Resource Description Framework (RDF) Model and Syntax Specification,* 22 February 1999 (http://www.w3.org/TR/1999/REC-rdf-syntax-19990222)

5.1 The AIP

The XML Schema for the AIP is the general document model in which the other XML Schemas can be used as subschema's.

The first AIP subelement contains the unique identifier of the AIP. The metadata and the various representations of the record form the content of the second subelement ('record'). The metadata relating to the record only has to be stated once and is stored in the element *preservation description information*. In the XML Schema the possibility is provided for mapping these metadata to the general metadata schema for electronic records, or in its absence, immediately to ISAD(G). Then the various representations of the record are included. In addition to the original representation, the XML Schema provides the possibility of including no, one or more migrated bitstreams. Nor does the original representation have to be present. This representation is not included, for example, for databases, analogue sources for digitised material, etc. For each representation, the relevant technical metadata is stated (*representation metadata*). The digital objects can contain binary files or XML documents. In cases where XML is used as the archiving format, a separate XML schema is used for these document models (see below).

The third child element of the root element 'AIP' is reserved for the registration of a checksum[22]. This checksum is calculated on the complete second child element 'record' of the 'AIP' including all XML tags and subelements of the subelements. The function of the checksum is to indicate the bit integrity of the AIP. Each bit change in the AIP will of course produce a different checksum and will be exposed by a comparison of the embedded and the recalculated checksum. Although in the OAIS reference model the *fixity information* is only intended for the *content information* of the AIP, in the implementation for the City of Antwerp the checksum is calculated on the complete 'record' element. There are two reasons for this. First, the checksum can then check the bit integrity of the embedded metadata and packaging information. This expands the function of the fixity information. Second, this requires less programming and this work method is easier to implement.

In future transformations (for example, when modifying the XML structure or adding a new migrated version) the checksum must be recalculated. If necessary, one can then use a stronger hashing algorithm because algorithms are becoming more vulnerable as computer power increases.

The XML schema for an AIP is available on: http://www.edavid.be/xmlschemas/aip.xsd

5.2 A record-keeping metadataset

Ideally, every archival institution or service has a general record-keeping metadataset for the management of electronic records. However, pending a general accepted standard there are just a few organisations with such metadataset for the moment.

Such a metadataset does not have to be built from scratch. It's recommended to use ISAD(G) as basis for the archival descriptions. The metadataset of the Premis working group is a good starting point for the technical metadata about the representations of an electronic records.

The presented record-keeping metadataschema is a work in progress. In the meanwhile some metadatasets are foreseen for an eletronic record, a file or subject folder, series, fileformats and

[22] The encapsulation of the checksum is technically not the easiest solution: one must be careful that the checksum is calculated on the correct bitstream and that components such as the XML epilogue, the XML tags for the <AIP> element and the complete element <checksum> are not calculated along with it. An easier solution is to preserve the MD5-checksum externally or to use it as the filename for the AIP container. In this last case, the semantics of the filename is lost, which makes searching for documents without finding aids almost impossible.

preservation media. The metadata elements used for archival description are linked to the corresponding ISAD(G)-elements by adding a reference to the ISAD-element numbers as attribute.

This record-keeping management schema is available at: http://www.edavid.be/xmlschemas/rkms.xsd

5.3 The archival descriptive metadata

Pending the development of a general metadata schema for electronic records, the City of Antwerp will include the archival descriptive metadata directly in the XML-AIP in conformity with ISAD(G). Since they relate to all possible representations of the same record, this description only has to be included once.

The XML schema for ISAD(G) has been designed in such a way that there is a choice between composing a complete ISAD(G) description and between the use of the individual descriptive elements of the ISAD(G) standard. In the first scenario, at least the six obligatory ISAD(G) fields must be included in the descriptive element. Since this is not always possible or needed, one can, instead of using a complete ISAD(G) descriptive index, also include individual ISAD(G) fields in this descriptive element. Much depends on the extent to which one can automatically compose the descriptive fields.

In the XML schema for ISAD(G) all elements have as an attribute a reference to the paragraph number in the ISAD(G) standard. Although the ISAD(G) standard prescribes that one does not *have* to use these numbers to designate descriptive elements, it is recommended for the sake of clarity and interchangeability.

The XML schema for ISAD(G) is available on: http://www.edavid.be/xmlschemas/isad.xsd

5.4 E-mails, calendars and databases

The City Archives of Antwerp uses XML as the archiving format for several types of electronic records:
- e-mails
- Outlook calendars
- databases

For each document type, an XML schema has been worked out. These XML Schemas can be used in two different ways: as the archiving file format or within the XML container files. The root element for each document type is “<document>” since the essential archival descriptive and contextualising metadata in the XML Schemas for the document types are lacking. This particular metadata information is part of the AIP container.

The XML Schemas are accessible on:
- e-mail: http://www.edavid.be/xmlschemas/email.xsd
- Outlook calendars: http://www.edavid.be/xmlschemas/calendar.xsd
- databases: http://www.edavid.be/xmlschemas/database.xsd

6. COMPOSING THE XML-AIP'S

Before the electronic records are ingested in the digital repository, the electronic records must be transformed into AIP's. Depending on the internal protocols and responsibilities, this transformation can be carried out by the creator and/or the archivist. The composing of AIP's involves several actions:
- migration of the original formats to suitable archiving formats
- encapsulation of the original and migrated bitstreams in XML
- registration and encapsulation of the essential technical and archival descriptive metadata
- generation of a checksum to check the bit integrity
- checking the quality of the XML-AIP's.

The origin of the embedded metadata can be divers. Preferably, the metadata should be recorded automatically as much as possible on the creation or receipt of the records. The metadata can be automatically extracted from the information system in which the electronic records are managed or can be fetched from the electronic records themselves. The condition for this is of course that these metadata are stored in a structured way. On the other hand, the metadata can also be assigned by the creator or archivist using a migration or encapsulation tool.

For the practical implementation of composing AIP's many different scenarios are possible. An important item that must be taken into consideration is the availability of software. It is best for records to be migrated to a suitable archiving format before the original software is no longer being used. And finally, encapsulation occurs for ingestion in the digital repository. These actions can be spread in time. The creator and the archivist can work out a distribution of tasks. One possibility is that the creator registers the metadata and migrates the records, while the archivist takes care of the encapsulation and adds any other metadata. Another possibility is for all actions to be carried out at the same time. When the creator composes the container files and transfers them in that form to the archivist, the XML containers then function not only as AIP's, but also as Submission Information Packages (SIP's). One can easily automate this if one has a migration or encapsulation tool for this, or equips the RMA with an AIP export module[23]. When an organisation, for example, switches to a higher RMA-version or another RMA-software, it is probable that not all records will be included in the systemupgrade or -migration. For that scenario, XML-AIP's can be exported from the old RMA.

Although migration and encapsulation can occur at the same time, this does not have to be the case. The City Archives of Antwerp developped a stand-alone migration and encapsulation tool for the creation of XML-AIP's. At present, this tool can be used to migrate and embed e-mails and word-processing documents completely automatically. As much as possible, the encapsulated metadata are automatically captured by fetching them from the information system or from the documents themselves. For the archiving of e-mails, XML-API's are composed that, in addition to the metadata, also contain the bitstream of the MS Outlook message format (original format) and of the XML version (suitable archiving format). For word-processing documents, in addition to the MS Word format, a multipage TIFF version is also embedded. Shortly, this tool will be expanded with a module for the archiving of MS Access databases.

7. CONCLUSION

Building further upon the digital presentation strategy recommended by the DAVID-project, a sustainable storage method for electronic records was developped. A solution is found in the combination of the OAIS-informationmodel with the physical encapsulation of all components of electronic records in one computerfile or AIP.

[23] Such an AIP export module can be used, not only for archiving purposes, but also for transferring the electronic records from one RMA to another.

The encapsulation of the various representations and the essential metadata in a container file is for many reasons an advantageous storage method. The different representations of electronic records are packed together with their metadata in containers with a view as it were to transport through time. Especially the advantage that the essential metadata are inextricably connected with the records is very important. This avoids risks, not only now, but also in future. The digital objects in the digital repository have increased autonomy and immediately have the status of record without being dependent on external information. The archival function or the recordness of the records is therefore preserved at all times.

The preservation of electronic records enriched with metadata also offers advantages for the digital repository itself. Electronic records form the basic units of the digital repository, are better protected against calamities and are better suited for future migrations and transformations. The digital repository requires little or no intelligence for the management of the electronic records, which enables the archivist to place less demands on the software/infrastructure for the digital repository. In principle, any robust storage system can suffice as a digital repository. All essential information is found in the XML-AIP's. This means that one also has little to fear from the technological ageing of the information system that is being used as a digital repository. One of the few functionalities of the digital repository is the random checking of the bit integrity of the preserved AIP's and the associated error handling when a problem occurs.

Automated archival finding aids can be part of the digital repository, but one can manage them just as well in related information systems that function as a finding aid. In the finding aids, the archival descriptions are linked to the ID's of the records in the digital repository. The archival descriptions can be composed partially automatically by extracting the metadata at the time of ingestion and/or description. The (embedded) metadata can also be indexed.

The City of Antwerp decided on the implementation of this archiving method in order to embed all representations (original and migrated formats) and the essential metadata in one XML container. This method does require that electronic records be transformed into container files on ingest in the digital repository. These container files are to a great extent self-descriptive, which is helpful for their interpretation. A good interpretation does assume that the archivist is familiar with the OAIS standard.