UN/CEFACT - ICA
eArchiving
Digital Record Exchange Standard
Jurisdictional Survey

1 Introduction

The eArchiving project aims at defining a standard for exchanges between agencies and archives, in order to facilitate automatic transfers of records or archives (more information is available on http://www.cen.eu/UNcEFACTforum/TBG/tbg19.htm).

This project was approved as a UN/CEFACT project in October 2006. It is supported by ICA (International Council on Archives) too.

As a first step of the eArchiving project, the purpose of this document is to elicit jurisdictional requirements for the exchange of digital records. Typically, this exchange would be a transfer from an agency (record creator) to an archive, but the transfer could apply between agencies, or between agencies and secondary storage suppliers.

It is based on a document prepared by the Australasian Digital Recordkeeping Initiative (ADRI).

1.1 Who is concerned?

This survey is intended for archives, records management services in agencies, storage service suppliers, related software vendors, bodies involved in eGovernment policies.

1.2 Structure of survey

In order to prompt responses, we have prepared a number of questions. For each question, we have prepared a short background explanation to assist you in understanding the issues that underlie the question.

Do not be concerned if you cannot answer, or are not concerned, with some of the questions. We would prefer to get a response with carefully thought out answers to the two or three questions that are really relevant to you, than anodyne answers to all questions. If we have few responses to particular questions, this is a good indication that the underlying topic is non controversial.

If, in completing the survey, you have issues that are not covered by the survey, we would love you to raise these issues.

1.3 When would we like your response?

We would be very grateful if you would provide feedback by the 15 January 2007.

1.4 Who do you send you responses to?

Please email your response to:

andrew.waugh@dvc.vic.gov.au

1.5 In which language should be your response?

Your response should preferably be in English. If this is not possible, do not hesitate to contact us.

2 Survey questions

In this questionnaire we will use the following terminology:

- Digital Record Exchange Standard (or DRES) when we are referring to the standard as a whole
- Submission Information Package (SIP) when we are referring to the package of information that is transmitted from the sender (e.g. agency) to the receiver (e.g. archive)
- Object when we are referring to a collection of information about something. Typical objects are: records and agents.

2.1 External references

- What external standards, codes of practices, or activities that you are aware of that will affect the development of the DRES?

Background

A Digital Record Exchange Standard will not stand in isolation, but will fit within a broader recordkeeping and archival framework. This framework may consist of international standards (e.g. the ISO standard on Metadata, OAIS) or codes of practice.

2.2 Simplicity vs Power

- Is it more important to you to have a Digital Record Exchange Standard (DRES) that is simple to implement, or one that offers flexibility to suit a range of conditions?

Background

A simple standard will be easier to design, document, implement, test compliance for, and maintain. It is also likely that implementations will be more consistent in their implementation, thus assisting interoperability between agencies and archives. On the other hand, simplicity (e.g. a restricted metadata set, or restrictions on how records may be structured) may mean less flexibility to implement a solution that suits a particular jurisdiction, agency, or system.

2.3 Maximum size of content

- What is the distribution of sizes of the content that you envisage transferring in a SIP?
- Do you envisage needing to transfer many large pieces of content (e.g. film or video segments)?

Background

The anticipated content size may affect the implementation of the DRES. For example, the transfer of very large content (e.g. films) may cause problems in fitting the transfer onto pieces of transfer media. This would require the ability to break the content into chunks and place the chunks in separate objects within the transfer.

The size of content becomes more of an issue if the content is binary and is to be included within an XML document. This is because an XML document cannot include binary data directly. Typically, the binary data is encoded as text using Base64, which results in slightly over a 33% increase in size of the content. (Every 3 bytes of the input file becomes 4 bytes, encoded, an additional 1 or 2 bytes (end of line characters) every 72 output bytes, and up to two padding bytes at the end of the content).

2.4 Transactions

- The DRES will cover the transfer of custody of electronic records between an agency and an archive. What other interactions is it important for the standard to support (e.g. destruction approval).

Background

While it is assumed that the DRES will be used to transfer records from agencies to an archive, the DRES could be generalised to support other transactions involving agencies, archives, and possibly other actors. These include for example:

- Destruction of records (either initiated by the agency or the archive, and possibly involving a third party authorising the disposal)
- Submission of a classification scheme by an agency to an archive
- Delivery of records to an end-user

2.5 Use

- Is it important to be able to use the Digital Record Exchange Standard to transfer records between other entities than just agency and an archive?

Background

While it is assumed that the DRES will be used to transfer records from agencies to an archive, the DRES could be generalised to support other transfers of records. These include:

- Archive to Agency (i.e. return of records to agencies)
- Agency to Agency (e.g. when functions are transferred between agencies)
- Agency to Secondary Storage (e.g. temporary records that are no longer of operational use and are to be stored off site)
- Secondary Storage to Agency (e.g. the return of temporary records stored off site).

2.6 Transfer process

- Should the Digital Record Exchange Standard include processes to ensure the reliable transfer of data between the agency and the archive?

Background

A Submission Information Package is simply a data package. Ensuring that this data package is transferred correctly between the sender and receiver requires a process to be defined. This process will need to be embedded in a system, probably with programmatic support. The process might include:

- Transmission of parts of the SIP separately and ensuring correct re-assembly at the archive.
- Detection of errors in transmission and recovery from the error (e.g. by retransmission).
- Return of application level acknowledgements (e.g. that the archive has accepted custody of the application).

Specifying this process in the DRES will add to the complexity of the standard, and limit flexibility. On the other hand, it will increase the potential benefits of the standard to agencies and archives.

2.7 Requirements on underlying transport protocols

- What functions should the DRES assume from the underlying transport or exchange protocol? For example, can it be assumed that the underlying protocol will ensure accurate transfer between actors?

Background

A DRES takes advantage of a underlying infrastructure to transfer objects between actors. How much should a DRES assume about the underlying infrastructure (and hence not be required to implement itself). Typical underlying functions would include: security infrastructure (identifying actors and ensuring error transmission); and the actual transfer mechanism.

2.8 Manifest

- Should the Digital Record Exchange Standard define a manifest that lists all of the records that are being transferred, or should it be a local matter between the sender and the receiver whether a manifest is used, and, if so, the form of the manifest?
- If a manifest is included, what information (metadata) should be included in the manifest?
- If a manifest is included in the standard, should the records be separate physical objects in the transfer, or should the records be contained within the manifest?

Background

A transfer will normally include multiple records. This raises the question of how to indicate to the archive which records are included in the scope of the transfer.

The easiest way is to delimit the scope is to depend on the surrounding process to recover from any errors. For example, the transfer is considered to be all the objects copied onto the set of transfer media, and the sender eventually carries out a reconciliation against the acknowledgement to detect objects that have been lost. (This is simpler than including a manifest as this reconciliation must always occur, even if a manifest is included, in order to detect any other transfer failures).

An alternative is to include a manifest (or list of objects) in the transfer. This manifest lists or contains all of the objects that form part of the transfer. The primary advantage of including a manifest is that the archive can immediately detect any errors – either objects that are supposed to be in the transfer but are missing, or objects that are not supposed to be in the transfer but are present. Secondary advantages of a manifest include:

- It is a place where information about the transfer itself can be represented (e.g. the identifier used to distinguish this transfer from other transfers)
- It allows the transfer as a whole to be examined without opening or manipulating the objects. (e.g. an archivist can scan the whole transfer to check that it is what was agreed to be transferred without processing each object in the transfer).

While these benefits are significant, it is not obvious that they are sufficiently important to require a manifest to be part of a DRES.

If a manifest is to be used, there are further questions.

First, should the manifest be standardised as part of the DRES? If the manifest was part of a local agreement between the sender and the receiver, this would simplify the standard. However, it would make implementation more expensive as code may not be shareable between jurisdictions.

Second, what information should be included in the manifest? Clearly, the manifest should include information about the transfer itself, and either the objects that form the

transfer, or references to them. The manifest could also duplicate a small amount of metadata about each object (e.g. title and classification), to assist the receiver in performing quality control on the transfer.

Third, should the records included in the transfer be contained within the manifest, or can they be physically separate objects referenced by the manifest? If they are contained within the manifest, then this requires that the whole transfer is physically one object. This makes it unlikely that the records could be separated and lost. However, it does mean that the manifest is likely to be a physically large object (potentially very large). This may have serious consequences for processing the manifest at the receiver.

2.9 Structure of transfers

- What are the conceptual models of the information (records) being transferred that need to be reflected in the standard?

Background

In developing a standard for the transfer of records, it is necessary to have a model of the information being transferred.

A record documents a transaction. Depending on the definition of a ‘transaction’, the documentation may be represented by multiple individual documents. These documents may or may not be arranged within the record (e.g. they may be arranged sequentially, hierarchically, or they may have no order what-so-ever). Each individual document may be formed from separate computer files (which together form the document); for example a database might be represented by individual files for the tables and schema. Alternatively, one document may be represented multiple times by different representations (e.g. a Word representation or a PDF representation).

Traditionally, records are grouped into files, and this conceptual model is followed in most (all?) ERM systems. A conceptual grouping into files is not necessarily followed in EDM systems, or other digital management systems, and so may not be a mandatory feature of digital records. (Alternatively, a ‘file’ may have to be artificially imposed on records originating from such systems.) Another way of looking at this is that a ‘File’ is a ‘transaction’ at another level, which contains a sequence of sub-transactions ‘Records’, which may, in turn, contain sub-sub-transactions.

Again, traditionally, files are arranged in some sequence and are accessed by indices or registers (often known as ‘control’ records). In a digital system, there are usually multiple methods of arranging and accessing records. Again, these may or may not correspond with traditional views of records. In an extreme case, there may be no static arrangement or index to the records. Instead, lists of ‘relevant’ files may be generated dynamically by searching. More typically, the files are arranged in a classification hierarchy (or there may be several classification hierarchies). Unlike conventional classification hierarchies, files may be found at other levels of the hierarchy than just the leaves.

Beyond the pragmatic arrangement of information described in the preceding paragraphs, there are also other entities represented (explicitly or implicitly) in recordkeeping systems. The SPIRT model, for example, contains entities that represent records, agents/people, business recordkeeping, business, and mandates. These entities all contain metadata, and the full context of a record is described by the interrelationships between these entities.

All of the entities identified in this section can be represented as individual objects in the transfer, or they could be combined into a single object. For example, there could be individual objects in the transfer representing an agent, or a document. Alternatively, there could be an object ‘record’ which could contain the information from the agent entity or document entity. Incorporating the information in the record objects is less

efficient (as the information is potentially duplicated many times). On the other hand this ensures that the context associated with the entity is grounded. If the context is included by reference to external objects, a decision needs to be taken on when changes to the external object are significant enough to represent a new object. For example, say a record is created that references an agent. Subsequently some metadata in the agent is modified, and then a second record is created. Can both records reference the same agent object? Is the answer to this dependent on what metadata has been modified? For example, a new occupant of the role might indicate a substantive change in the agent, but not a change to the telephone number.

The final consideration is whether the objects in a transfer need be related. A transfer could be restricted to contain only related objects; for example, the objects related to a particular series. At the other extreme, a transfer could be a completely random collection of objects from disparate series.

2.10 Support for the Transfer of Physical Objects

- Should the DRES include support for the reference of physical objects (e.g. paper files) in addition to the transfer of electronic records?

Background

Agencies will continue to generate paper records for a significant period of times. Files in some series may be paper based. It is also expected that some files will have a mixture of paper and electronic records. It is quite possible that an electronic file may logically include paper records. Should the DRES support either joint paper/electronic files, or series in which some files are paper?

2.11 Transfer Objects

- Should objects include content and metadata, or can external objects and metadata be referenced?

Background

Each transfer will consist of a number of objects representing entities. Entities consist of metadata and content. A basic decision is the whether the metadata and content must be included within the object at the time of transfer, or may be referenced externally.

Using an external reference for metadata or contents allows considerable flexibility. For example, the reference could be a URL, or a database query. Referencing content avoids problems with the maximum size of the object, and means that it is not necessary to encode binary data for inclusion in an XML document.

Physically including the metadata and object within the object fixes the object and ensures that it stands alone. Fixing the object ensures that the object has not changed between the time the object was created and when the content was retrieved. For example, if the object represented an agent/person, the metadata associated with the agent could change dramatically over time, or even be deleted. Fixing the object means that the receiver knows exactly what the object is.

2.12 Object identifiers

- What form of object identifiers should the SIP support?
- How unique should they be?

Background

Every object in the transfer needs to be uniquely identified – for some level of uniqueness and for some period.

Normally, objects are uniquely identified within the context of the recordkeeping system in which they are held. When transferred to another system they need to be uniquely identified within that system, which may involve a change of identifiers. The identifiers could be changed before they are sent, or after they are received.

If the identifiers are changed, or at least qualified, by the sender, then the objects are always uniquely identified, even when received by the receiver. The qualifier would identify the record system which originated the object. Typically, the record system would be identified by the agency and series identifiers. If desired, the archive can be uniquely identified to make the full identifier globally unique.

When the receivers are changed, or qualified, by the receiver, then objects are being delivered to the receiver that are probably not uniquely identified (as two record systems may assign the same record identifier to two different records).

If the identifiers for the archive, agency, and series are appropriately chosen, the full identifier can be persistent (i.e. should never need to change).

2.13 Metadata

- Should the Digital Record Exchange Standard standardise the metadata about objects?
- If it should standardise the metadata, should the DRES reference an external standard (if so, which one), or explicitly define the metadata?
- If it should not standardise the metadata, should the DRES provide holes in which any metadata collection could be inserted?
- If the metadata is not standardised, should the DRES standardise a small number of metadata elements that are used for managing the transfer? If so, what metadata elements?

Background

Apart from the content itself, all of the information associated with a record is represented as metadata. For some objects (e.g. Files, Agents), the object consists entirely of metadata. There are a large number of metadata standards available for describing objects (e.g. PREMIS, SPIRT)

The first question that needs to be considered is whether a standard DRES needs to specify a standard set of metadata at all. It would be quite possible for the DRES to specify an object skeleton that contains slots into which any metadata could be fitted. This approach has both advantages and disadvantages.

The advantage is that this avoids having to standardise metadata within the DRES standard, making the standard flexible and much simpler to define. This is a particularly attractive approach because different jurisdictions will have different requirements on recordkeeping metadata. Further, it is clear that international metadata development has not yet been completed.

There are corresponding disadvantages as well. The most significant is that not setting a metadata standard significantly limits the benefits that can be derived from a DRES. For an archive, a major benefit of a DRES is that it provides a common agreed set of metadata. The mapping from the agency specific metadata to a common standard is undertaken at the agency before the object is sent. This is the best place to perform the mapping as the agency has the most information about their specific use of metadata. Mapping into a standard metadata set is simpler and cheaper for product vendors as well, as it provides a single fixed metadata set. Without a standard metadata set, vendors are likely to have to implement several metadata mappings when they sell products in different jurisdictions.

A second disadvantage of not having a standard metadata set is that it may be difficult to write generic software to process DRES objects. For example, the unique identifier of an object is a critical piece of metadata which is used in processing an object. If this unique identifier is not in a standard place, then it will be difficult to extract.

It would not, of course, be necessary to specifically define the metadata in the DRES standard. If an external standard is suitably well defined, the DRES could simply refer to it. The disadvantage of this approach is that may become hostage to changes in the externally defined standard.

An intermediate position between not specifying any metadata (and leaving this to implementation), and fully specifying the metadata (either in the DRES, or by referencing a metadata standard), is to define a small number of key metadata elements in the DRES. The standard will also include 'holes' into which richer metadata can be slotted. Examples of metadata that might be standardised in this way include:

- Unique identifier
- Relationship
- Title
- Classification
- History
- Access control
- Technical information about the data format of the content

This advantage of this approach is that the software dealing with transfers can obtain the information necessary to process the transfer without needing to process the remaining non standard metadata.

2.14 Authentication

- Is it necessary to seal the SIP to ensure that it has not been tampered with?
- Is it necessary to cover the entire SIP by the seal (i.e. the manifest and all objects that form part of the transfer), or is it sufficient to cover only part of the transfer?
- Can one seal be used to cover the entire SIP (even when the transfer is physical split into separate objects), or should each object have its own seal?
- Are seals only used to protect data during transfer, or will they be required within an archive

Background

A seal (e.g. a digital signature) allows the detection of any corruption in the SIP.

The first issue to be considered is whether the SIP needs to be sealed at all. If the SIP is not sealed, the records are not directly protected from modification between export from a record system and receipt by an archive. (Note that they might be indirectly protected, for example by encrypted Internet traffic, or security procedures on a physical transfer).

If a SIP does need to be sealed, the next question is how much of the SIP should be covered by the seal. Clearly the content of the records should be covered by the seal, but should the context? The answer to this is almost certainly yes, as the context forms part of the record. Should the manifest be covered by the seal? If the manifest was not covered by a seal then records could be added or removed from the transfer without the possibility of immediate detection. However, a properly designed transfer process must be designed to detect the loss or addition of records outside any protection provided by the SIP. This is because the records may be lost (or added) before the SIP is constructed (i.e. in the source recordkeeping system), or after the SIP has been received by the archive.

If the objects in the transfer are separate objects to the manifest, should a single seal cover the entire transfer, or should separate seals be applied to each object? In theory

only one seal needs to be applied, and this would cover the manifest and all referenced objects. This approach is potentially more difficult to calculate (as the software calculating the seal has to switch objects at the correct point). Calculating separate seals on each individual object has the advantage that if an object is corrupted, the seal only breaks on that one object, not on the whole transfer. This greatly simplifies the complexity and cost of recovery (and also may assist is diagnosing the problem).

Will the seal only be used to protect the SIP during the transfer, or will they be retained in the archive to allow subsequent revalidation of the objects? One view of seals is that they protect the records while they are outside the control of a system (either a recordkeeping system or archival system), where they are vulnerable. While they are being held within a system, the system itself is relied upon to detect prevent or detect corruption. Typically, this would be by means of a hash value (checksum) to protect the integrity of the content, and security mechanisms built into the system to protect the metadata (including the hash value). An alternative view is that the seal should be retained by the archive to allow the eventual end user to verify that the object has not been changed since creation by the original agency. This allows end users to treat the archive as a 'black box', and reduces the level of trust required in the whole system. It, however, increases the complexity of the system. It requires a mechanism to allow an archive to modify an object while retaining the validity of the original seals. It also requires a mechanism to ensure that sufficient information is available to verify the seals while the records are held by the archive.

2.15 Definition of standard components

- Should the standard define components that are re-usable in other business protocols?

Background

Many business protocols are considered in several fields. For example the exchange of health data between health centres or the exchange of custom data between different countries. These protocols may also need to convey some archival information (e.g. retention period, accessibility...). Should the DRES provide core components (information models) usable by such protocols?