Electronic Records: Problem Solved?

Justine Heazlewood, Jon Dell'Oro,
Leon Harari, Brendan Hills,
Nick Leask, Ainslie Sefton,
Andrew Waugh and Ross Wilkinson

A Report on the Public Record Office Victoria's Electronic Records Strategy

Justine Heazlewood, Public Record Office Victoria; Jon Dell'Oro, CSIRO; Leon Harari, Ernst & Young; Brendan Hills, CSIRO; Nick Leask, Ernst & Young; Ainslie Sefton, Public Record Office Victoria; Andrew Waugh, CSIRO; and Ross Wilkinson, CSIRO.

In 1998 Public Record Office Victoria (PROV) initiated a project to examine the capture and long term preservation of the electronic records of the Victorian Government. The project team, consisting of computer scientists from CSIRO, business process analysts from Ernst & Young, and archivists from PROV, built a system which successfully demonstrated that it is possible to capture electronic records with existing technology and from existing systems in such a way that electronic records can be preserved in the long term.

The Electronic Records 'Problem'

Archival and record keeping systems to manage paper records have been part of government for centuries and have been developed to a sophisticated degree. Today, however, government business is becoming electronic. Currently government is at an interim stage where most government documents are created electronically, but are printed and managed as paper documents for exchange and archiving. The growing use of email and the Internet is pointing to a future where government activity is conducted completely electronically.

There are a number of reasons why electronic government activity should be recorded electronically rather than on paper.

- *The evidential status of a paper form of an electronic record is unclear.* An electronic record documenting a transaction which has been carried out electronically, may lose evidential weight if copied to paper, as it may not represent the 'original' record.
- *The cost of storing electronic records is much lower than that of storing paper records.* The decreasing cost of electronic storage combined with the increasing cost for paper storage means that it is no longer cost-effective to store all records in paper form.
- *A paper record of an electronic transaction may not capture all aspects of that transaction.* Printing electronic records to paper may not fully document the context of those records.
- *The cost of finding records is much lower and searching capability is greatly extended in an electronic environment.* Unless paper records are well managed and well documented, they can be difficult to find. An electronic environment allows more sophisticated ways of accessing and retrieving records.

The creation and management of records produced in an environment where desktop computers have replaced pen and paper has been of concern to both records managers and archivists for some time. There are a number of reasons for such widespread concern:

- *Document formats change and become unreadable.* For example, it is nearly certain that Microsoft Word documents that are created today will not be readable in 100 years time.
- *Electronic objects can be subject to undetectable change.* Unless precautions are taken, it is possible for an electronic file to be altered without any way of detecting that a change has occurred. The evidentiary status of that record may be compromised because it can be undetectably modified.
- *Electronic records may not be captured because most record capture processes are paper based.* Increasingly email, e-commerce, and electronic documents are being used to conduct business, but the processes of record keeping, developed over a long period of time, are paper oriented. To date much record keeping practice has required that electronic records are printed and then incorporated into the existing paper based record keeping system. Unless a conscious decision to print the electronic record is made, the record will not be archived.

- *The context of an electronic record, and its relation to other records, can easily be lost.* In an electronic environment the context of record creation can easily be lost if it is not documented at the time of record creation. For example, the time that a document is emailed to a client may be crucial, but this information is not usually recorded as part of the document.
- *Capturing context can be expensive.* Context is crucial to the understanding of the record in the future, but it may be too expensive to realistically expect to capture all the appropriate contextual information manually or to add contextual information to the record at a later stage.
- *Existing systems for managing electronic documents and records are not designed as archive systems.* An archive system preserves the content, structure, context, and evidential integrity of the record for as long as the record is required. Existing electronic document and records management systems do not provide the functionality required in order to archive records for long term accessibility.

Archivists wish to ensure that records which document government policy, individuals rights and entitlements, and other classes of records identified as being of permanent value, can be managed in a way that ensures their continuing existence and accessibility to future generations. Archives and archivists have been dealing with the 'problem' of electronic records in different ways, but no archival agency has yet found a practical solution.

Keeping Electronic Records Forever: Identification of the Problem

Public Record Office Victoria (PROV) has initiated a series of projects designed to provide it with a new strategy for dealing with electronic records. In 1996 PROV was provided with funding from Victoria's Microeconomic Reform Fund to address issues of electronic records management and archiving. The project culminated in a report called *Keeping Electronic Records Forever.*[1]

The project was undertaken by Ernst & Young and the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) in conjunction with the PROV. As the project was designed to establish a conceptual answer to a technological problem, the report developed a business case that explained the recommended solution in conceptual terms and identified the key issues upon which it impacted. The project undertook an intensive document review and a series of workshops involving many of Victoria's larger agencies. Representatives from the Archives Authority

of NSW and the Australian Archives (National Archives of Australia) also participated.

The project team concluded that there was a fundamental need for a 'static' *record* that was inviolable and which satisfied the evidentiary requirements of courts and government. It further concluded that agencies required 'static' *records* to which they could refer for their own operational needs, but also the full functionality that came with computer software applications, in order to create, modify and manipulate *documents.*

Victorian Electronic Records Strategy: Solutions

In 1998 PROV began the Victorian Electronic Records Strategy project to implement the recommendations made in *Keeping Electronic Records Forever.* The project was run by the PROV, again, in conjunction with the CSIRO and Ernst & Young. The project developed a prototype system to demonstrate that it was possible to capture records from existing business systems in such a way that they would be accessible in the long term.[2]

Keeping Electronic Records Forever advocated that instead of taking a *system* oriented approach to electronic records, a *data* driven approach was more appropriate, as the records would outlast any system developed to manage them. In order to tackle this problem the project team undertook as series of investigations so that they could understand the government processes which led to records creation and how records were used and managed in the same context. The archival processes of PROV were also investigated. Using this information, they pinpointed the information required to be captured when creating an electronic record.

The second key point made in *Keeping Electronic Records Forever* was that electronic records should be captured at the time of creation. There are several strong reasons for this requirement. Firstly, the record is more reliable as evidence if captured at time of creation. Secondly, there is more chance that the record will in fact be captured if it is done immediately. Thirdly, information capture at the time the transaction is undertaken is both cheaper and more reliable than *post hoc* data entry.

The team developed several prototype electronic capture environments using the Victorian Department of Infrastructure as the source of records and record capture processes, to demonstrate that records could be effectively and cheaply captured in typical government processes. Attention was paid both to the capture of records where activities occurred in a comparatively *ad hoc* manner, and where the underlying process was very regular and amenable to conventional workflow solutions.

Based on PROV descriptions of its archiving activities, an archive system demonstrator was developed along with a retrieval system demonstrator. Each was built as a stand-alone system although it was recognised that in actual production systems there are a number of possible development strategies. The value of building these systems separately was that the team were able to demonstrate that the electronic record was self-contained and able to be passed between systems reliably, without further information.

Having built these systems, several sample working environments were created to demonstrate successful capture, archiving and discovery of electronic records. A series of demonstrations to PROV, government agencies, records managers, vendors and other interested parties were held which were able to both effectively demonstrate electronic archiving and also provided an opportunity for valuable feedback on the approach. This feedback was incorporated into the project and refined the VERS approach.

The VERS project was successfully completed in December 1998. PROV released the *Final Report* of the project in April 1999. Standards for Electronic Recordkeeping for the Victorian Government, which are based on recommendations made in the VERS *Final Report,* were also published at the same time. PROV are currently working on the implementation of compliant systems within the Victorian government and on establishing archiving systems to manage VERS style records in the PROV.

Electronic Records

Electronic records are simply the computerised versions of traditional paper records created and kept by agencies. Sources of electronic records range from desktop applications such as Word, Excel, and email, to corporate applications such as financial systems, HR systems and corporate databases. Typically records are evidence of government or organisational activities and include policy documents, memos and letters, and database reports. Theoretical archival science contains a number of formal definitions of 'electronic records' but, from a practical perspective, the project needed a definition which gave the widest possible scope to agencies and PROV to define records for their own purposes. A record, to an agency, is simply whatever information they need to preserve. Victorian agency decisions will always be directed by legislation, PROV standards and agency practice.

The project concentrated on the preservation of electronic documents. Documents comprise the majority of government records, and most documents are created electronically on desktop computers. The project

also decided that focusing on a manageable subset of records would be more fruitful than attempting to cover all types of electronic records.

The project defined two component parts of electronic records:

1. Content
Content is defined as the original information that is being preserved. There are many different types of content; for example documents, databases, and images. Content types may be encoded in different formats. The Long Term Electronic Record format is sufficiently flexible to contain a variety of content encodings and types.

2. Metadata
Metadata is defined as information associated with the content of a record. Metadata can describe the record, describe the content of the record, document its relationship with other records and the organisation (the record's context) and describe the encoding of the content.

There have been a number of metadata proposals for electronic records. Two well known North American proposals are *The Preservation of the Integrity of Electronic Records* (which derives from work carried out at the University of British Columbia),[3] and *Towards a Reference Model for Business Acceptable Communications* from which was derived the *Functional Requirements for Recordkeeping* promoted by the University of Pittsburgh.[4]

The structure of the recommended Long Term Electronic Record is based on, but not identical to, the Pittsburgh model. From an archival perspective, there is little to choose between the two approaches, however, the team found that the Pittsburgh model was far easier to utilise as the Pittsburgh metadata was more detailed and directly addressed the question of what specific pieces of information would be necessary to support computer based archiving of records.

The Pittsburgh model is a general model that can cover any type of record keeping system, however implementation of the full Pittsburgh model would impose significant costs, and many of the features were very specific to particular views of records and record keeping. To avoid excessive implementation costs, metadata fields that were not necessary for archiving in the Victorian government context were coalesced or eliminated. A full discussion of the eliminated fields can be found in the *Victorian Electronic Records Strategy Final Report.*[5] However, the VERS record structure has been designed so that the eliminated fields can be easily reinstated if required without affecting the functionality of the system.

The team also extended the Pittsburgh metadata to support additional metadata fields that were of specific interest in the Australian context such as the Australian Government Locator Service (AGLS) metadata. Some additional fields were added and some slightly rearranged to support the specific technology used.

Long Term Electronic Records

Electronic records which are able to be accessed in the long term need to meet the following criteria:

- *Long life.* Records must have an indefinite life. That is, the contents of a record must be capable of being viewed forever as the users originally saw them. Therefore the records must be in a form that can be physically preserved (for example easily copied from one media to another). It is also essential to preserve the indices and context of the record as to preserve the content itself so that records can be found. The context should be sufficient that the record is able to be understood.
- *Evidence.* A major reason for preserving records is to be able to legally prove what actions were taken and why they were taken. Electronic records must consequently be admissible as evidence and given due weight in a court of law. In practice this necessitates the ability to prove who created the record, when it was created, and that the record has not been subsequently altered.
- *Disposal.* Not all records need be preserved forever. Some records, indeed, must be destroyed after a period. Long Term Electronic Records must be able to be sentenced and destroyed if necessary.
- *Augmentation.* Not all information about a record is known when the record is created. A record may, for example, need to be refiled. It is necessary to be able to augment or change the information associated with a record without disturbing the evidentiary integrity of the record.

The VERS Long Term Electronic Record structure is expressed using XML (eXtensible Markup Language). XML is a text based mark-up language and is easily extensible (unlike HTML) and are relatively simple. The XML standard is defined in *Extensible Markup Language (XML) 1.0.* A number of document encodings can also be used within this structure. The VERS demonstrator encoded document content using Adobe System's Portable Document Format (PDF) Version 1.2.

Long Term Electronic Records must also be:

- *Self documenting.* It must be possible to interpret and understand the information in the record, at least at a primitive level, without reference to external documentation (which might have been lost). To this end the recommended structure:
    - Is based on ASCII text. This means that the structure of the record can be viewed using the most primitive of computer tools.
    - Contains short textual descriptions of any more complex encoding of information.
- *Self contained.* The electronic record structure must contain all information about a record. It is far easier and more reliable to manage the information associated with a record if it is stored in one place rather than in components which are stored separately.
- *Extensible.* It is simple to extend the structure to add new metadata or new record types without affecting interoperability of the recommended basic structure.

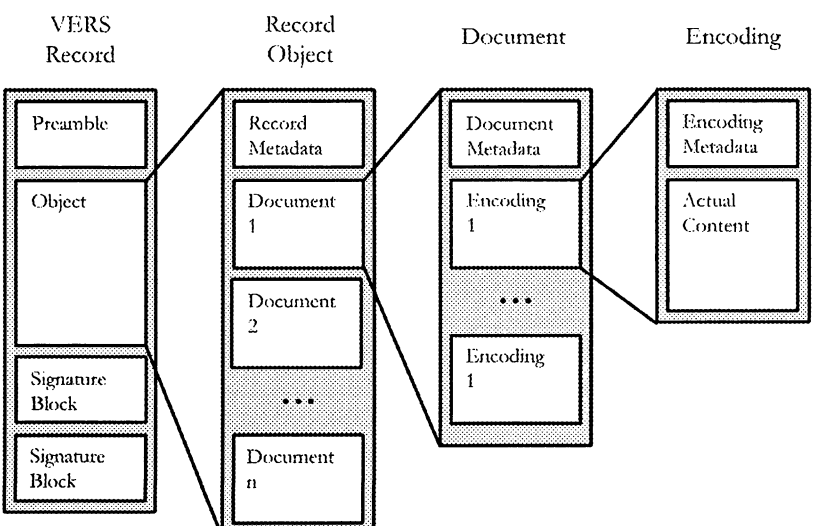

Figure 1: Standard Long Term Electronic Record Structure

The standard record structure is shown in Figure 1. A generic VERS record may contain many different types of object; in this case, it contains a record. A record contains metadata general to the entire record and one or

more documents. A document contains metadata specific to itself and one or more encodings of the document. An encoding contains a representation of the actual content and metadata that describes how the content was encoded for inclusion in the record.

The information in a VERS record includes a preamble, one or more signature blocks, and the object contained within the record. The Preamble describes the type, purpose, creation date, and basic format of the object in the VERS record.

Each signature block contains the result of signing of the record. The signature validates the preamble and the object content. The signature blocks are included for evidential reasons. It is necessary to be able to prove who created the record, when it was created, and that the record has not been subsequently altered. This is required not only to guard against forgery and alteration by the creators of documents, but also to guard against forgery and alteration by the administrators of the Archive system itself.

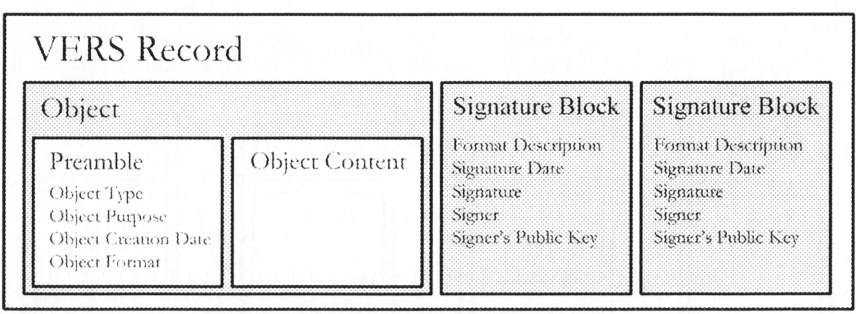

Figure 2: Generic VERS Record

The final component is the record object itself. A record has two parts. The first part is record level metadata that deals with the record as a whole and includes:

- *Handle information.* The unique identifier of this record and descriptive information.
- *Contextual information.* This includes information about the transaction documented by this record, who is responsible for the creation of the record, links to related records, and documentation as to the reliability of the system itself.

There are other fields which may be included in the record metadata. These are:

- *Terms and conditions.* This element contains information about the policies which affect the record. This includes who may access the record, how the record may be used, and disposal information.
- *Use history.* This element contains the history of the record.

The VERS solution recommends that this information be stored separately rather than in the record because it may be subject to rapid change or continual addition and would therefore compromise theinviolability of the record.

The second part of the record object contains the content of the record and may be one or more documents. Each separate document is represented in the record as a separate unit. A document may be a set of data (for example, a report, a CAD file, a database, Web pages, or datasets). Different types of document will be encoded differently and will need different metadata, but the Long Term Electronic Record format has a standard structure for documents which can be customised to handle different document types.

A document contains document level metadata, and one or more encodings of the document. The document level metadata contains descriptive information about the document (for example its purpose), and information about the source of the document (a textual description of the application that produced the document). For certain types of documents the source description may be very detailed and may need to be structured. For example, if the document was a database, the source description could include the database schema so as to aid interpretation of the data in the database. Another example would be a scientific data set (for example an image from a satellite). In this case, the data source would include information about the instrument that captured the dataset, and the instrument settings used.

The data that forms a document may be encoded in many ways when the document is included in the record. For example, a report could be encoded as a PDF file, or a Word file. A document in the record can contain several representations (encodings) of the same physical document.

In the VERS model, documents are represented using PDF. The primary selection criteria for a document format was confidence that, for the foreseeable future, it would be possible to write a viewer for the document from publicly available information. Word file format, for instance, would not be an appropriate format, as the description of this format has not been published. The PDF standard has been published and is freely available.

PDF is also flexible. PDF can be generated from any application that can generate Postscript (the standard printing language); thus anything that can be printed can be represented in PDF. PDF can also be generated from scanned documents. Scanned documents can be converted to an electronic document which is very close (or identical) in appearance to the original paper document. The text of the scanned image can be accessed, altered or used after employing optical character recognition software. PDF is reasonably efficient in terms of size and, in fact, much more efficient than Postscript. One disadvantage of PDF is that it is a binary format and hence must be encoded into text before inserting into the record. The VERS project used the Base64 standard to encode PDF. Base64 is an Internet standard that is a fundamental component of email systems. A second disadvantage of PDF is that there is nothing (except market pressure) to force Adobe to abide by their own published standard in the future. This opens the possibility of PDF files that cannot be viewed using a viewer based on the published PDF standard.

The VERS Demonstrator System

The VERS demonstrator system consisted of three major components:

- *Record Capture.* This component simulated the desktop environment of a government agency. Record capture was implemented by integrating different desktop applications in representative workflows and capturing records into the Long Term Electronic Record format.
- *Repository.* This component managed the archived records, including tracking the location of records, sentencing, and destruction.
- *Record Discovery.* This component allowed users to search for and display archived records. The functions included in this component were the building of indexes from information contained in the records, searching for records, and the display of retrieved records as Web documents.

Record Capture

Record capture is the process of creating electronic records and their metadata. It must capture both the content (the information contained in a record) and structure (how that information is organised).

From an archival perspective, it is important that both the content and structure are accurately captured. The captured record should be identical in appearance to the original document as it was viewed by the creator of the record. This requirement restricts the technologies that can be used to

capture documents. With HTML (Hypertext Markup Language, used for World Wide Web documents), for instance, the appearance of the document depends on the settings of the user's browser and the resulting record may bear no resemblance to the appearance of the original document.

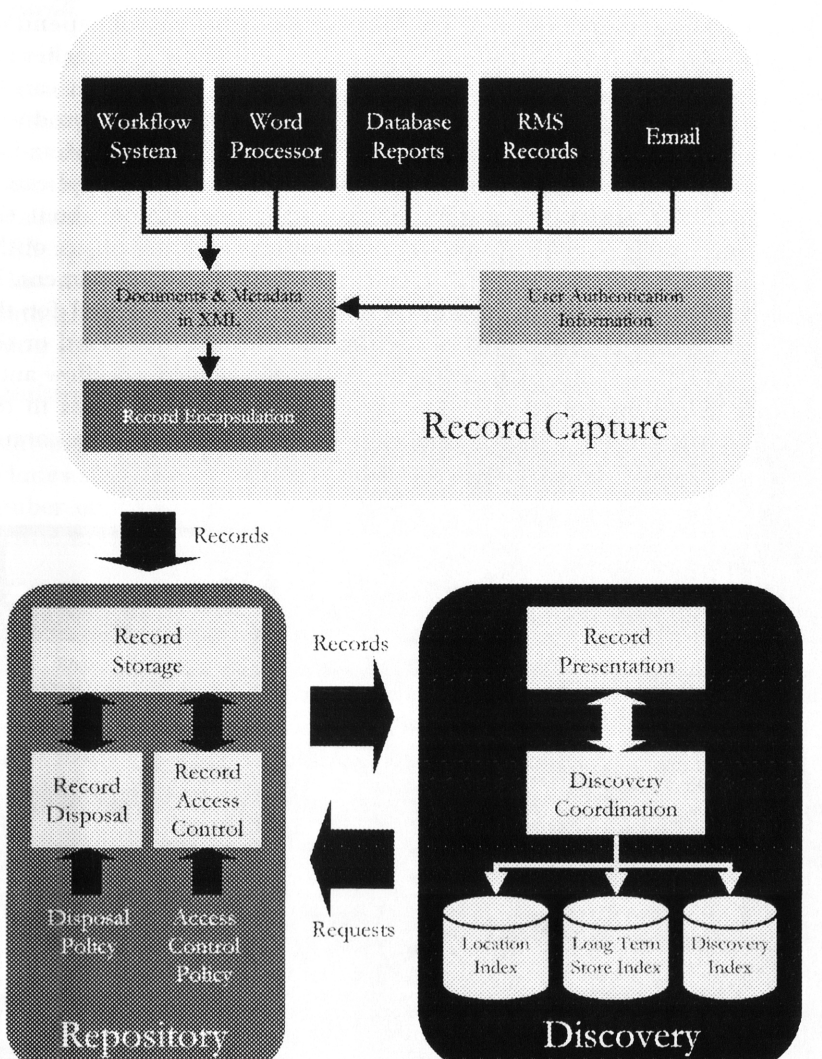

Figure 3: Demonstrator System Design

Metadata can be generated in a variety of ways. For instance, the record creator can enter it at the same time as the record is created. But full manual metadata entry is tedious and unlikely to be performed well. The VERS project determined that record capture systems could automatically generate much of the metadata thereby relinquishing the need for manual entry. The methods used to perform this task in any given system will depend on the particular application capturing the records, but many approaches are available. Metadata which depends on the record capture system can be automatically generated by that system. For example, all records produced by a particular process will have the same technical description. Metadata can be derived from the document itself, although this can be application dependent. For example, email headers contain information about the sender, the recipient, the time and date of sending, and the subject of the email. Metadata can also be derived from the computing environment. In particular, the record's creator and its time of creation (essential for the evidentiary status of the record) can be obtained from a smart card or the user login. Finally, a record generated from a programmed workflow automatically has a context (that is, a relationship with other records in the workflow). Metadata entered at one stage in the workflow can be carried along with the workflow and added to later records.

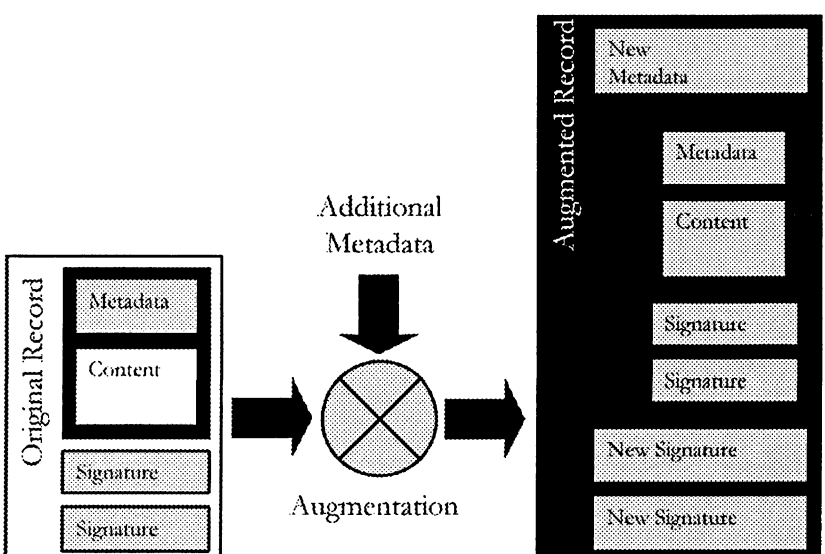

Figure 4: Augmented Record

The record capture component used in the VERS demonstrator system captured the content and context of a record. The content was captured using PDF, and the metadata was captured in an XML file along with the digital signatures which secured the record.

Augmenting Records

The digital signatures that secure a VERS record also prevent any modification of the record. In some circumstances, however, it might be necessary to modify the metadata associated with a record. For example, the need may arise to refile a record, to add additional descriptive information, or to make a new linkage to another record. To allow the metadata to be modified without disturbing the evidentiary integrity, the VERS approach allows an entire record to be included as the content within a new record. This layering of record metadata is referred to as creating 'Onion Records' (see Figure 4). The content in the case of an Onion Record is a complete record.

Repository

Archiving is the process of managing records over a long period and includes storing, preserving, and providing access to records. There are a number of issues which arise when these functions occur in an electronic environment.

Computer technology rapidly becomes obsolete. A CD, for example, may be readable in one hundred years, but it is unlikely that there will be any working CD reader at that time. Therefore it will be necessary to move records from one medium to another as storage technology changes. This process is often termed 'refreshing.' Refreshing may also be carried out to protect against record loss due to deteriorating media (for example a tape wearing out), or to make a physical copy of the record (for off site storage for instance). The process of refreshing can be made entirely automatic.

In an electronic Repository it is possible to rebuild, or even create new, indexes from the archived records. These indexes may be internal to the archive (for example the index that maps a record key to its physical location), or external (for example the indexes within the Discovery system that allow searching). Rebuilding allows these indexes to be constructed from scratch. In particular, it allows the indexes to be rebuilt after a catastrophic system failure. It also allows the commissioning of a new Repository or Discovery system (in this case the archived records are used to initialise the indexes in the new system).

Disposal practice may be affected by an electronic environment. In the Victorian government disposal is carried out at the file level. The destruction of a file implies the destruction of the records within it. Sentencing of records is based on disposal policy and these may change as the result of departmental or governmental determination. The actual policies will vary across agencies, but some disposal policies require storing of significant information. The VERS approach recommends storing information about actual disposal decisions in a file record (that is, a record which contains information about a group of records). The VERS Repository can automatically apply disposal policies to the records and construct a list of files scheduled for destruction. The application of sentencing should be a conscious decision by a records manager and not left to an automated system. The system can log the decision to dispose of records, including who authorized disposal, the date, and the disposal policy that caused the disposal.

In both agency and archival environments it is necessary to control access to records. Access control is a based on legislation and agency policy. It will vary between agencies and change over time. Because access control policies will change and may change rapidly, storing access control policies in the records will cause two problems when modifying the policy. First, it will be necessary to locate each record affected by the change in policy (there could be thousands). Second, it is necessary to modify the metadata of each of these records and this becomes unacceptable in a system which relies on the inviolability, and therefore inalterability, of records. Instead of storing the access control policy with the record, the recommended alternative is to store it in a separate database. This database can be amended as frequently as desired, and the policy is only 'activated' when a record is accessed.

Discovery

Being able to follow contextual relationships (like the link between incoming correspondence and outgoing replies) provides a powerful mechanism for finding information. Most of the contextual information used in the VERS demonstrator system was captured when the record was created and was stored as metadata within the record. The basic contextual relationship used in the Victorian government is collecting related records into files. To aid in management of records, particularly at the PROV, files are assigned to series (collections of related files) and series to agencies.

Officers who access files on a day to day basis will understand the purpose of individual files within the agency context. To assist other users and

researchers' understanding of the context of the records, archival finding aids are created to describe the history and function of the agency, as well as the purpose and organisation of the files and series within the agency. Finding aids can be stored within the Discovery system.

There are, however, other contextual relationships which can be documented. One contextual mechanism is the thread of transactions. A transaction is an interaction such as sending a letter. By linking related transactions together (for example the incoming letter to the authorisations to the response sent) it is possible to follow the chains of cause and effect.

Ad hoc relationships are essentially random links between related files or records and may be created for any purpose at any time. A typical *ad hoc* relationship is a linkage between versions of a record. *Ad hoc* relationships can be documented in a record's metadata but the user creating an *ad hoc* relationship must consciously document that relationship.

One-off relationships may be established by the researcher. One way of doing this is by full text search, that is searching on the words contained within the record. Full text search may be carried out on the metadata associated with the record, or the content of the record (or both).

The Discovery System used in the VERS demonstrator provided the user interface to the archived records. It consisted of a set of database tables which stored metadata in XML, documents in standard electronic record format, and record linkages.

The VERS demonstrator made use of Web based technologies and delivered records via a Web browser. The desktop costs of Web delivery are negligible. All modern desktop machines have a Web client built in (Internet Explorer) and alternative clients (for example Netscape Navigator) are freely available. Web clients have extensive built in software for displaying most data formats (for example, PDF).

As the archived records were kept in a database, and had extensive metadata associated with them, it was possible to construct a much more powerful searching environment than with paper based records. The Discovery System allowed:

- *Searching via a finding aid and original indexes.* This reflected current practice within PROV, where records are found by searching the original agency indexes. PROV also develops finding aids to describe the purpose and structure of the agency and the files within the agency.

- *Searching the content of the metadata.* Searches were able to be performed on any combination of the metadata fields.
- *Searching the content of documents within the record.* The VERS demonstrator system stored the text for each record in the database so that it was possible, for example, to return all records that contained a specific word.
- *Searching on any combination of the content of the documents and metadata.*
- *Searching on record context.* The VERS demonstrator allowed records from a particular transaction to be linked together.
- *Searching on File context.* All records were associated with a File, and a group of records in the same File were able to be retrieved.

Conclusion

The role of the records manager becomes increasingly important in an electronic environment. The context of individual records and files can be captured using the VERS model, but more global information about the organisation as a whole could easily be lost. Records managers are needed to ensure that information about the *way* that an organisation keeps records, as well as the records about the organisation as a whole, survive the move to a digital world.

The VERS project has successfully demonstrated that the capture and long term preservation of electronic records is possible now. The VERS project offers a solution to the electronic records 'problem' which has plagued the record keeping profession for some time. Further work needs to be undertaken to deal with very complex electronic records, but the VERS approach gives us the ability to tackle the clear majority of the electronic records being created now.

We 'know' the past by the records people have left. Future generations, who may well have 'known' the late twentieth century as silence and absence, will now be able to judge us by the many voices we leave behind.

Full details of the VERS project can be found in the Victorian Electronic Records Strategy Final Report available from Public Record Office Victoria or at http://www.vicnet.net.au/~provic/vers/.

Endnotes

1. *Keeping Electronic Records Forever: Records Management Vision Development,* Public Record Office and Ernst & Young, 1996. Available at http://www.vicnet.net.au/~provic/vers/kerf.htm

2. The project team took, as a putative figure, the period of 100 years to represent the 'long term.'

3. *The Preservation of the Integrity of Electronic Records,* School of Library, Archival and Information Studies, University of British Columbia, 1997.

4. *Towards a Reference Model for Business Acceptable Communications,* David Bearman, December 1994.

5. *Victorian Electronic Records Strategy Final Report,* Public Record Office Victoria, April 1999.