gla.erm/erm/ermproj/2001-02-21/8

"No Going Back?"

The final report of the Effective Records Management Project

Project funded under the JISC Technology Applications Programme.
JTAP - 375

James Currall, Claire E. Johnson, Pete Johnston, Michael S. Moss, Lesley M. Richmond

ISBN 8 5261 759 3

Acknowledgements

University of Glasgow Information Strategy Working Group

JISC JTAP especially Tish Roberts and Tom Franklin

Arthur Allison

Neil Beagrie

Julie Cargill

Colin Farrow

Peter Ford

Peter Kemp (original co-grant holder)

Jane Lee

Neil Leitch

Linda McCormick

Dugald Mackie

Lynne Moss

Tony Prosser

Seamus Ross

Alec Scrimgeour

Ellen Stewart

Alistair Tough

All the people who attended the seminars and offered advice.

Contents

Executive Summary

This report covers both project methodology and exploration of underlying intellectual issues of improved information management in the digital order.

Record creation, management and long-term storage in the digital order requires a diverse skill set drawn from a number of disciplines, which have not traditionally communicated with each other:- administrator, computer scientist, archivist, librarian and information services manager. Attention needs to be paid to record keeping in the digital order because:-

- An increasing proportion of records in Universities are now produced digitally (albeit with printing in mind),
- More flexible distribution and consultation methods are possible,
- There is a need for a reduction in paper storage,
- A high level of 'consistency' is desirable,
- A better integration of record keeping with other aspects of process can now be achieved.

The requirements for an effective digital record system are developed in Chapter 4. The overall objective of the ERM project was the provision of 'protocols and tools for the effective management of information in the digital order, with particular attention to information held in a document-based form'[1]. The project developed a demonstrator system which addressed the full range of issues involved in the management of a 'testbed' of digital records in the form of documents, from their creation through distribution and use to their final disposal or permanent retention. The testbed selected was the records of a subset of the university's committees and was used to develop ideas and to assess how well they worked in practice. This testbed system and the way in which it fits into the Glasgow Committee system is discussed in Chapter 5, with additional technical details in Appendix 2.

A brief outline of the relevant issues in archiving and records management is provided in Chapter 2 and the key challenges presented by the digital order and discussed in chapter 3 with further development in Appendix 4 are:-

- Digital documents can only be viewed using computer hardware and software, which is constantly changing, placing the documents at risk. This issue is discussed in Chapter 3 and the steps that need to be taken to minimise this risk discussed in Chapter 7.
- The media that digital documents are stored on degrades over a relatively short time-scale (a small number of years), placing the documents at risk. This issue is developed and the steps that need to be taken to minimise this risk discussed in Chapter 7.
- Traditional markers within documents that allow reference to be made to particular parts of the document (such as page numbers) are not necessarily reliable in the digital world, as different programs, browsers, printers, etc. render the same document differently. This issue is discussed in Chapter 3.
- Digital documents may be moved from one computing environment to another. It is necessary to be able to refer unambiguously to a particular document wherever it is and not to spend a large amount of time dealing with name clashes as documents are moved around. This issue is discussed in Chapters 3 and 4 and the solutions developed during the project are outlined in Chapter 7.
- Digital documents are much more easily tampered with than paper ones. Measures need to be adopted which reduce the likelihood of falsification of digital records. This issue and the approaches to solving it are discussed in Chapter 8.

The ERM project wished to exploit the benefits of reusability which accrue from a structured approach to document creation. It was recognised, however, that it was not feasible to demand

[1] Effective Records Management Project, *Project aims and deliverables*, (University of Glasgow ERM Project, April 1998). http://www.gla.ac.uk/InfoStrat/ERM/Docs/deliv.htm

that a large number of record creators of varying skills and backgrounds should adopt the use of a new set of software tools specifically for this purpose. Our approach has been to make it easy for document creators to meet the requirement of the developing digital world, by enhancing the tools that they already use, rather than requiring them to learn to use new ones. A variety of different approaches to this problem are evaluated in Chapter 6.

In documents on which the ERM project has concentrated, making the structure explicit is a first step to the documents adding value to the processes of which they are part. A document in the digital order can and should be used for further processing within the transactions of which it forms a part, rather than simply being a step in the creation of a representation of information on paper. In the paper order documents do not have this potential. If advantage is not taken of these possibilities, we are simply using a more expensive set of tools to create the documents that we previously created with typewriters, with few advantages to show for the additional expense.

The project ensures the following benefits:

- clearer identification of the roles and responsibilities for key records creators
- improved understanding of record creators about the creation, use, dissemination and disposal of digital records
- training of users in good practice of digital record management
- improved retrieval speed for documents and elements within them
- increased accuracy for items within a document and the individual document within a collection
- reduction of organisational risk from unmanaged records
- appropriate retention of records and better resource management

This project report indicates that the ERM project provides an investment in the future, opportunities for better information use or re-use, identification of legal risks and identification of good practice whether you be an administrator, a computing scientist, an archivist, a librarian or an information services manager.

1. Introduction

1.1 Prologue

Until the beginning of the twentieth century, documents created as records were treated as single objects, and referenced in registers, sometimes with multiple entries. The First World War multiplied the information flow and made speedy retrieval critical, which lead to the introduction of the file, where related documents were collected together sequentially. The drawback of the file is that documents, even if they do not relate to more than one subject, have to be filed in more than one place. For example, letters relating to more than one subject have to be filed under each topic.

This replication was made possible by technological change – the development of the typewriter and means of copying documents.

1.2 Surrogacy

Copying documents raised issues of the difference between originals and surrogates. Where documents were deemed to be 'significant', for example a treaty, a legal agreement, private ledgers and so on, they were accorded special protection to ensure that the 'original' or authentic copy could always be located. This was achieved by locking them up, sometimes in containers where two or more keyholders had to be present before retrieval was possible. In some institutions the system of retaining the originals as the record of a transaction persisted, even if copies were filed elsewhere. This reflected the fact that information contained therein related largely to one transaction.

In most file series, originals along with surrogates were filed without discrimination, although references were often attributed to indicate the author and creator (if internal) or the provenance (if external). Instead of each document being registered, the file became the unit of reference. Like the system they replaced, files were on the whole managed by third parties (registries mostly, but also secretaries). The integrity of this fiduciary model was guaranteed by the fact that the authors of the information mostly used a third party to create the document, secretaries, clerks and so on. It would be wrong to suggest that the file entirely replaced the old system of recording each document as a single unit. In some organisations this system persists or persisted until recently.

1.3 Practices and procedures in the paper order

Over the centuries conventions and protocols became embedded in the paper order. For example, it is easy to distinguish between a letter and a memorandum or between a formal letter and an informal letter, both from their language and content (vocabulary and structure). For example a personal letter may be signed 'love Basil' whereas a formal letter may be signed 'yours sincerely, Basil Bulstrode, Professor of Ancient History'. From the user's perspective there was value in these conventions if they addressed the problem of provenance, authenticity and verification and facilitated retrieval. These elements were documented by archivists in their practices and procedures. This became the study of *Diplomatics*, to test authenticity and veracity, and became their stock in trade. Names for groups of documents or elements within a document acquired precise meanings that were understood, albeit implicitly.

It was known that the act of transferring a document to a registry or an archive guaranteed survival, retrieval and strengthened authenticity in a way that placing it in a desk drawer did not. In other words it became endowed with evidential value.

1.4 Migration to the digital order

These systems had begun to break down by the 1990s. The arrival of the word-processor blurred the distinction between author, creator and editor without any consideration being given to the consequences for authentication, verification and retrieval. With the advent of networked

PCs in the late 1990s, many author-creators began to distribute information digitally, usually as a means of reallocating costs – saving on their photocopying budgets and the tedium of filling envelopes, with little thought to filing, referencing and future use.

This practice often leads to frustration. The document may be unopenable because platforms across user communities are hardly ever uniform. If a document can be opened and printed, pagination is infrequently the same between print-outs. Since there is rarely consistent internal referencing (section numbering) in the paper order, this hampers committee work. Users are naturally irritated by such difficulties and either revert to the paper order or simply turn up to meetings ill-prepared.

Even more confusing was the unthinking of paper terminology by the digital but without its precision of meaning. The word 'file' is used in the digital order to mean what the Americans quaintly but accurately call a 'bunch of data', without necessarily implying any common factor or effective means of retrieval. A file can be 'archived', which means transferring it to an external disc or tape that has an uncertain life and does not endow it with evidential value. Files can also be placed in digital directories that bear no resemblance to the directories that are the well structured finding aids of the paper order. The digital order's use of this vocabulary has led to much misunderstanding between information technologists and information professionals.

1.5 The digital revolution?

The situation at Glasgow when the ERM project began work resembled that described in section 1.4. The team brought together practitioners drawn from a variety of backgrounds, to ensure that concepts of practice and procedure were fully understood. It was recognised that the issue was not one of the enabling technology, but of the information itself. In some senses new technology is not new but just a further step in a process which started when writing began. Such an evolutionary viewpoint reduces the risk of claiming too much for the technology, the cardinal sin of those who see it as a revolution and talk grandly of an 'information society'. Today there are those who advocate that all records should be created and distributed digitally. At the most extreme the apostles of the 'new age' propose retro-conversion from paper to digital[2] on a massive and wholly unaffordable scale. This is at present an unworkable option, as are suggestions that all information will be read from the screen. The idea of snuggling up in bed with a portable is laughable. The two orders, like those of handwriting and printing, will endure.

Such an understanding has important consequences in the design of protocols and procedures for the digital order. The users' need for paper outputs, which are consistent one with another, remains paramount. However if all the digital order provides is a more convenient postal system then it will be condemned as a massive waste of money.

Compared to the paper order it is expensive to maintain the digital order. The digital order must add value; ease of retrieval, navigability, and record linkage are all much trumpeted as powerful features of even the most basic software. This is perhaps best characterised as the 'you can do things quicker' syndrome.

1.6 Process

The intellectual management of information in the digital order is addressed by several international projects (for example InterPARES led by Luciana Duranti[3]). The premise that information should be preserved for reference for as long as it has value remains unchallenged, but the questions that are still largely unanswered are those which relate to who is to be entrusted to evaluate, manage, protect and ensure access.

---

[2] *A Meta-Evaluation of Electronic Document Management Systems* University of Manchester (August 2000). http://www.man.ac.uk/intra/subproj/proj17.htm

[3] with Heather MacNeil, 'The Preservation of the Integrity of Electronic Records: an Overview of the UBC-MAS Research Project,' Archivaria 42 (Spring 1997): 46-67.
'The Concepts of Reliability and Authenticity and Their Implications,' Archivaria 39 (1995): 5-10.

As discussed, the file emerged about a century ago through convergence of technological innovations, the typewriter, the duplication process and the resulting growing volume of paper. Its introduction was informed by and informed changes in the way business was transacted. At present there is a similar combination of imperatives which is linked to the question of the legal admissibility of records which may not be consistent with the medium on which it was created[4]. This may then conflict with which media provides the most durable form of long-term preservation. The requirements of the Data Protection Act (1998) mean that while some records may be preserved for historical use, elements within them may have to be deleted or erased within a specific, shorter, timescale. For the archivist and records manager this legal or compliance environment reinforces the continuing need for their services in the digital order to determine the evidential value of a record and ensure appropriate retention/ disposal.

Records management systems need to be customised to meet the current and future needs of the organisation. In all sectors the document environment has changed so dramatically that attention has been diverted from such concerns by the overwhelming volume of documents in the digital and the accompanying problems of poor security and long-term preservation of the media. Choices in technology make the preservation of records more complex. Specifically it means record creators as well as archivists have to make an assessment about what constitutes the complete file and how to integrate its disparate parts across media. It cannot be expected that all the documents will reside in one medium alone. When this is coupled with a poor general understanding of archival or preservation requirements, it is not surprising that information is being negligently lost. Within active record systems vast quantities of paper and digital documents are being kept unnecessarily often in a confused filing structure that does not assist access, let alone the process of appraisal and preservation. The penalties for not managing records properly range from increased administrative costs resulting from inefficient working practices to serious legal liability and a damaged reputation, for example mis-filing records containing defamatory text or revealing information given in confidence.

1.8 Compliance

As the ERM project was being conceived there were increasing external constraints and concerns affecting the management of information by organisations that had their origins in consumer protection. Already data protection legislation was in place, but this only related to personal information held in electronic form and the penalties were relatively innocuous. By the time the project began work the current government had signalled its intention of signing the European conventions on human rights and much tougher data protection legislation embracing both the paper and the digital orders was under consideration by the EC. Consumer expectations were also changing rapidly in the wake of pension mis-selling and other scandals. This required all organisations to be much more disciplined about retaining records of the process underlying transactions, for example in *Higher Education Institutions* (HEIs) procedures for admissions, examinations, and classification of degrees. This required a review of systems for managing such records in the digital order.

As the project progressed it became increasingly clear that the evolving compliance regime, as it came to be known, would have a major impact on the information landscape. It demanded that many of the practices for managing information that have been carelessly abandoned in the migration to the digital order needed to be reintroduced. The risk of being non-compliant both in terms of reputation and of penalties far outweighed the increased associated costs. The problem for HEIs and other government bodies was that unlike the private sector and even *Non-Governmental Organisations* (NGOs) these costs could not be passed on to the customer through the price mechanism. The consequence for the project was that practices and procedures had to be devised which could be accommodated, as far as possible, within existing staffing structures.

Compliance also had the effect of concentrating attention much more firmly on issues of surrogacy. This is a particular problem in HEIs where often for very good reasons additional information is added to a record with little regard to the core (or original data), for example, information about class prizes is added to student records held in departments but not the record

[4] , see for example, 'Recordkeeping and electronic mail policy: the state of thought and the state of practice' David A Wallace 1998.

held by Registry. Moreover a record can itself be a compilation, being made up of surrogates of other records, where, typically in an HEI, the decision making and ratification process are embedded in complex hierarchies. A good example is the design of new courses by Boards of Studies and their subsequent ratification by Faculties and finally Senate. Although it is well known that the original may be located in a subsidiary record, this is rarely transparent as the cross-referencing may not be adequate. This makes hard-wired linkages[5] in the digital order almost impossible without a change of culture and practices. Moreover compliance demands much more attention to the effective disposal of records in the digital medium.

1.9 Project management and philosophy

As a result of all these pressures the ERM project had to address from the outset a much more complex set of questions than had been originally anticipated. The project was based in the University Archives, the traditional custodian of records in the paper order. The team comprised Michael Moss, professor of archival studies, Dr. James Currall, user services manager in the Computing Service, Lesley Richmond, acting university archivist, Claire Johnson, senior records manager, and Pete Johnston, the document analyst and programmer. They were supported by two records managers and later ancillary technical staff with programming and training responsibilities. The experience of developing the JISC funded Information Strategy in 1997 as one of ten pilot sites had reinforced the view that in an organisation as large and complex as the University of Glasgow, the development and implementation of any system would of necessity have to be incremental. Consequently it was decided to concentrate on committee papers and to adopt an 'exemplar' approach using the committee system of Information Services (of which the University Archives is a part) as a 'testbed'. As practices and procedures were developed and proven, it was anticipated that other parts of the university would actively seek to adopt them for their own committees and this is what happened.

This report reviews the issues which the team addressed and sets out to explain how they can be answered. Inevitably the discussion is complex and at times theoretical. The key issues that re-appear fall within the following themes:

Technologies

How new theories and applications of information technology are shaping and reshaping our information preferences and expectations, information systems and services.

Knowledge

How discourse communities, fields of knowledge and information ecologies are defining and redefining themselves in changing technological, social and political contexts.

People

How individuals, and their many diverse communities, interact with their information environments, both technological and intellectual, at a cognitive, cultural or intellectual level.

These three themes interact in a complex web that is not, at present, fully revealed.

[5] This is where a direct hypertext link is made between sites or specific elements.

2. Records and Archive Management

2.1 Introduction

Records and archive management have their own precise terminology which has been hijacked by the information technology profession. The words '*record*' and '*file*' have a different meaning in each domain. Data can be 'archived' in an IT system in a way that is not recognised by a records or archive management system, so much so, that the latter system would identify that the data had not been 'archived'. The ERM project team, consisting of information professionals from a variety of backgrounds, had to ensure in discussion that concepts of practices and procedures were fully understood by their colleagues in order to avoid misunderstandings. The processes and purposes of records and archive management and clarification of associated terminology are described in this chapter.

2.2 Records management

2.2.1 Records management and record definitions

Records management is an activity established by an organisation to achieve economy, efficiency and effectiveness in the **creation, distribution, use, maintenance, storage, and disposition** of all types of records created or received by that organisation in the course of its business[6]. These functions enable processes within the system to achieve the objective of maximising information benefits while minimising related costs. In essence a records management system provides the right information, to the right person at the right time, for the right length of time, at the lowest possible cost.

A *record* is a piece of recorded evidence or information, created or received by an organisation or person for use in the course of business and subsequently kept as evidence of such business. Records are created to communicate decisions, to provide evidence and to collect and process information. A record has certain characteristics. It forms part of the activity that creates it and is evidence of that activity. It is related to other records that perform the same function or contain similar information (see 2.3.2). It has permanence; an action that changes it creates a new record.

Because records are related one to another they are stored with systems which recognise such fundamental relationships[7]. Alphanumeric documents are normally held in files organised by subject and with contents in date order. Accounting transactions are nowadays stored in databases that allow them to be aggregated according to a number of typologies. The underlying paper transactions, if they exist, are stored sequentially and referenced in the database. The file is a recognisable entity within a records management system and is constructed in accordance with an over-arching file plan or a scheme of classification. However all systems are of necessity dynamic, reflecting the changing needs and functions of the owning institution.

Active records are ones that are used on a regular basis; *semi-active records* are ones used on an irregular basis; and inactive records are ones that are rarely used but are still required as compliance or fiduciary evidence.

2.2.2 The purpose of records management

Records management systems are put in place to provide cost-effective creation, exploitation and access to the records of an organisation. It enables cost reduction or containment, legal protection, disaster recovery, efficient storage, effective retrieval, and maximise information use and value. Records management will provide competitive advantage, improved customer

[6] See BS ISO 15489-1:2001 for *Information and documentation - Records management*
[7] See Jay Kennedy and Cherryl Schauder *Records management a guide to corporate record keeping,* 2nd ed. 1998.

service, improved staff morale, standardisation, auditable procedures, compliance and accountability.

2.2.3 Records management functions

2.2.3.1 Creation

There is a cost associated with the creation of any record, particularly if there is a requirement for it to be retained for audit. A transaction in a shop will produce a receipt for the customer, but it will also be recorded internally to show that certain items have been sold and therefore stock reduced, and perhaps that VAT has been charged. Customs and Excise officials and auditors may wish to examine it in the future, this requires the shop to have in place the necessary processes. In a small shop these may be nothing more than a sales book and a shoe box. In a multi-national chain there will be an elaborate IT infrastructure to hold the information. Some documents are more expensive to create than others, as more time and effort is invested in them, for example contracts and examination results. Even in small organisations, the 'value' of such documents is often recognised at their point of creation, but may subsequently be overlooked. A records management system is designed to improve the effectiveness of the record and enhance the future manageability of the record.

2.2.3.2 Distribution

Almost all records are created to communicate information and therefore a records management system should ensure that records are distributed in a cost-effective manner to appropriate recipients. Recipients ought to have confidence that they have no custodial responsibility as the authentic original is held for safekeeping within the records management system.

2.2.3.3 Use

Records are used to support an organisation's business processes, from decision making to compliance. A records management system must ensure the effective retrieval of the correct records to the appropriate person and ensure that those with no right of access are unable to access the records.

2.2.3.4 Maintenance

Records are at their most active immediately after their creation. Thereafter they require to be maintained for varying lengths of time. A records management system establishes the period of time that records are to be maintained, organises them in such a way that retrieval is effective and efficient, and provides protection to ensure that the integrity and authenticity of the record is maintained.

2.2.3.5 Storage

All records need to be stored in an environment that ensures protection from natural, technical or human damage or interference[8]. Appropriate storage methods may differ for active records

[8] There are a variety of projects addressing a variety of contemporary issues including the development of national and international standards. One such project which addresses the need of electronic records-keeping systems is the 'Model Requirements for the Management of Electronic Records' (MoReq) specification which is available at http://www.ISPO.cec.be/ida and http://www.dlmforum.eu.org. In Australia the 'Victorian Electronic Records Strategy' (VERS) available at http://www.prov.vic.gov.au/vers/final/finaltoc.htm a broader remit and developed a test bed system to prototype a "future state" electronic document processing and record capture system, using the Department of Infrastructure as the source of records and record capture processes. The prototype enabled techniques for dealing with electronic record creation, management and archiving to be demonstrated, put on trial and evaluated.

and inactive records, as retrieval requirements are not equally acute. Digital records require the same protection in specialist facilities, the requirements for these are discussed in chapter 8.

2.2.3.6 Disposition

Few records of an organisation, estimated at only 2-10 per cent, have enduring value and are retained permanently for the lifetime of the organisation; the remainder have a finite life. A records management system establishes the period of time that records are to be maintained, this is codified in a retention schedule. The system will also ensure the timely and secure disposition of records at the end of their life. The length of time a record must be retained is often stipulated by external authorities, for example the Inland Revenue, the Customs & Excise, and a variety of regulatory bodies. Because of their commitment to scholarly endeavour and the subsequent academic interest in their outputs, *Higher Education Institutions* (HEIs) have tended to retain more records permanently than other organisations.

2.3 Archive management

2.3.1 Definition

Archive management is an administrative procedural system established by an organisation to achieve economy, efficiency and effectiveness in the **selection, maintenance, preservation, access** and **use** of that portion of the records that has been selected or is designated for permanent preservation. There is increasingly a conflict between archival and records management systems, records management is designed to ensure that records are destroyed when they are time expired in compliance with external constraints, whereas archival systems are designed to ensure permanent preservation. Both systems, however, share a common characteristic in the retention of information in a fiduciary environment, which guarantees long term authenticity. Since documents that are known to 'exist' can be discovered by judicial process, many organisations are reluctant to preserve any records permanently except those they are required to by law, such as formal records of certain types of transactions. Paradoxically the recent public interest in the whereabouts of the assets of holocaust victims has encouraged a longer term view, as those institutions which have maintained archives are able to respond to specific enquiries.

2.3.2 Archive management systems

Archive management systems are put in place in order to provide cost-effective preservation of records for access by public or private communities, in a fiduciary environment to ensure probity and veracity of the records. They ensure the intellectual ownership of the records is understandable to the accessing community, managing them to preserve their information content, physical integrity and authenticity. Systems enable efficient storage and effective retrieval of records selected for permanent preservation and maximise information use and value.

2.3.3 Archive management functions

2.3.3.1 Selection

Records are selected for permanent preservation compliant with external constraints. An archive management system negotiates and accepts appropriate records from creators/custodians. The system identifies the designated accessing communities and determines the level of understanding required to be provided. In the past creators or owners have often applied selection criteria with little consistency or objectivity. Information is retained because it is thought to be useful or interesting, but in the process much information of possibly historical significance has been destroyed. (In this archival context historical does not just mean of interest to historians – narrowly defined – but to anyone who has need to understand past events.)

The capability of the digital order to store vast accumulations of information has on the surface removed the imperative for creators or owners to be selective. The danger remains that records are not managed and may be retained when legally they should have been destroyed[9]. Similarly *Personal Computers* (PCs) may be discarded while still containing sensitive records. This is avoided by having effective archival/records management policies in place. Some records can be designated for permanent preservation at creation, such as minutes of meetings and significant contract documents. The challenge is to encourage a more deliberate approach to selection.

2.3.3.2 Maintenance

An archive management system arranges records according to the principles of provenance, original order and function, in such a way that retrieval is effective and efficient. The system manages records to preserve their information content and authenticity. Strictly applied, such principles often retained random groupings of papers because they had been found together.

2.3.3.3 Presentation

Archive management systems must ensure that records selected for permanent preservation are understandable to the accessing community. Metadata should follow international descriptive standards and cover such areas as provenance, form and informational content of the material being described, as well as information about the creator and function. Finding aids describe, control and provide access to the archive material. Access to individual records is provided through points of access such as creator, date, subject, or unique id.

2.3.3.4 Access and Use

Archive management systems make the preserved information available to designate communities. A system supports users in determining the existence, description, location and availability of information stored in the archive and allows users to request and receive records.

[9] The Office of the Information Commissioner (http://www.dataprotection.gov.uk/dpr/dpdoc.nsf) provides the most obvious guidance on this.

3 The Challenge of Digital Documents

3.1 Introduction

The overall objective of the ERM project was the provision of 'tools and protocols for the effective management of information in the digital order, with particular attention to information held in a document-based form'[10]. The principal deliverable was a demonstrator system which addressed the full range of issues involved in the management of a 'testbed' of digital records in the form of documents, from their creation through distribution and use to their final disposal or permanent retention. The testbed selected was the records of a subset of the university's committees.

The first part of this chapter surveys some general features of the digital environment and suggests that the management of a digital record require different techniques from that of the paper record. The second part focuses more closely on characteristics of digital documents. For a fuller treatment of the facets of the digital document see Appendix 4.

3.2 The digital record

A record is a piece of information generated or collected in the course of an activity and kept as evidence of that activity. Any unit of information goes through a life-cycle in which it is created, used in some way (including, perhaps, in ways its creator did not anticipate) and either placed in storage or disposed of. Some information is of value only for the duration of a specific, short-term task and can be destroyed once that task is complete; other classes of information may have a legal value that dictates that they are retained beyond the period of immediate 'active' use. Historical permanency extends the record's life beyond the lifetime of the organisation imbuing it with an historical value that dictates permanent retention. The patterns of access and use of the same unit of information may vary greatly across the different phases of its life.

Various agencies, both internal and external to the institution that performed the activity and created a record, have the right to request that this evidence is produced in the future. There is an increasing range of regulations, conventions, and precedents for how long such agencies can expect records to be retained. For example the Local Government (Scotland) Act 1973 stipulates the Council meeting papers and reports should be available for a period of six years. The Information Commissioner's Code of Practice for CCTV specifies the appropriate retention period for images (Fifth Data Protection Principle) and once the retention period has expired, the images should be removed or erased (Fifth Data Protection Principle)[11].

As in most other organisations of comparable size and complexity, much of the information resource of *Higher Education Institutions* (HEIs) is now created, stored, distributed and used in digital forms as a matter of course, and a proportion of this information may constitute records.

The records manager or archivist has the responsibilities of identifying and classifying records, assessing their long-term value, and ensuring that they are disposed of or retained as appropriate, including measures to support their preservation and accessibility, regardless of the medium of their storage and use.

[10] Effective Records Management Project, *Project aims and deliverables*, (University of Glasgow ERM Project, April 1998). http://www.gla.ac.uk/InfoStrat/ERM/Docs/deliv.htm

[11] For example, publicans may need to keep recorded images for no longer than seven days because they will be aware of any incident such as a fight occurring on their premises within that time. Images recorded by equipment covering town centres and streets may not need to be retained for longer than 31 days unless they are required for evidential purposes in legal proceedings. http://www.dataprotection.gov.uk/dpr/dpdoc.nsf

3.3 The digital document

The ERM project has concentrated on the management of records in **document** form.

Establishing a clear definition of 'what is a document' is far from trivial. We recognise a 'paper document' largely from its physical structure and the medium on which it is stored. It is impossible to pass off a letter written with a fountain pen on paper from wood pulp as a medieval document, except as a surrogate and even then the reader would need evidence of the whereabouts of the original and perhaps even a facsimile to be confident about authenticity. In the digital domain, however, all resources are stored as bits on a wide range of optical or magnetic media. Consequently knowledge of physical structure and storage medium does nothing to help us establish whether a resource is a document or not. Even a focus on a document as 'textual' is problematic, since it does not satisfactorily encompass images and other objects or even various representational forms.

Buckland[12] argues that although concerns with documents in digital forms may have generated renewed interest in the question, it is one which has for a long time exercised those traditionally concerned with managing printed and manuscript resources. Given that the different technologies used to create, store and use the objects which are recognised as 'documents' have such different characteristics, definitions based on form and medium are inherently limited. More useful, Buckland suggests, is a 'functional approach' which accepts that what is considered to be a document will vary in these different environments, but that those objects will perform a recognisable role, which is that of supplying, in material form, evidence of a physical or abstract phenomenon. Levy[13] adopts a similar position with his view that documents are things created to 'speak' on behalf of human beings, with the expectation that their message will be reliably carried through space and through time.

Buckland's emphasis on evidence seems quite close to the definition of a 'record' (see 2.2), although a records manager may argue that within a typical organisation there are a large number of documents which do not have value as records, except in a most ephemeral sense. The most obvious and most quoted example is receipts and invoices, although arguably these are of interest to certain constituencies. However the Records Manager unlike the Archivist has to justify retention on the grounds of utility to the organisation. In this case, a distinction is being made on the basis of the 'phenomenon' of which the object is supplying evidence. On balance the interest of the Records Manager focuses on records that provide quality evidence for the activities of the organisation.

The ERM project, however, recognises that it has not addressed the management of all those digital objects that might fulfil such criteria. There has been a necessary pragmatic element to the scope of its work, which rests on a contingent notion of the document. It has, therefore, focused on those digital objects produced (principally) by word-processing and text-editing tools.

Even within this narrower domain, however, there are cases that the project has not addressed:

- a digital document may be a **transient unit** assembled from a set of component parts which are themselves liable to change subsequently - for example, data values extracted from a spreadsheet or database, or an assembly of parts of other documents which are themselves in a state of change. Such documents are ubiquitous in HEIs with their complex hierarchy of checks and balances embedded within their committee structures. To quote the example of the Board of Studies again, even in the paper world elements

[12] Buckland, Michael, 'What is a 'document'?', *Journal of the American Society for Information Science* 48(9): 804-809 (September 1997). Draft version available at
Buckland, Michael, 'What is a 'digital document'?', *Document Numerique* 2: 221-230 (1998). Draft version available at http://www.sims.berkeley.edu/~buckland/digdoc.html

[13] Levy, David M., 'Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment', in *Authenticity and Integrity in the Digital Environment* (Council on Library and Information Resources, May 2000). Available at http://www.clir.org/pubs/abstract/pub92abst.html

from minutes were cut and pasted into the subsequent papers and minutes of Faculties and Senate.

- the **extent** of a paper record is, in part at least, defined by the boundaries of its physical structure, or the boundaries of a physical unit of which it is part. Increasingly, the use of hypertext linking means that digital records are **not bounded** in the same way: they contain (and may be contained within) a complex set of relationships with other resources. The sense in which these other resources are an intrinsic part of the record may be difficult to define. Although this might be seen as analogous to the presence of external references in the printed record, it might be argued that the capacity to **traverse** a hypertext link, to 'de-reference' the target, represents a significant component of the user experience of the record. Everyone has experienced the frustration of using links to urls which no longer exist or to pages where the information has been removed because it is no longer considered current. Preserving such a capacity, however, may require considerable resources: ultimately, the decision on where a record's boundaries lie, on what proportion of this (potentially vast) web of resources constitutes a significant part of the record, is an arbitrary one. Although the same problems may exist in the paper order, the external references are not dependent on such transient objects as hypertext links but on a range of finding aids with well-developed protocols and procedures for access, distribution and preservation. For example, a book quoted in the footnote can be found in the library catalogue, located and read.

The project's work on the demonstrator system has concentrated on those records where it is possible to capture a static bounded representation.

Although, at the time of writing, there is no work which formally develops the idea of what properties or facets of a document-based record might constitute a Lynchian 'canonical form', there have been various efforts to identify in less formal terms what essential properties must be preserved. For example, the DLM-Forum's *Guidelines on best practices for using electronic information* identifies four facets of a digital record which are significant in conveying information:

- content
- structure
- context
- presentation

and it stipulates that the first three of these facets must be preserved.[14]

3.4 Lessons

The use of digital documents which at face value seems simple and straightforward is much more complex and requires input from creators, records managers and IT professionals. It, in no sense, replaces effective records management and is not, in itself, a solution to the problems of bulk and retrieval. To be accepted digital record keeping and distribution must provide added value to the user in terms of access and retrieval. It can only be implemented as a component of a considered institutional strategy which observes agreed conventions. Simply pressing a button to save as html has no obvious utility unless such a strategy is in place.

[14] DLM-Forum, *Guidelines on best practices for using electronic information* (Office for Official Publications of the European Communities, 1997)
http://www.ispo.cec.be/dlm/documents/guidelines.html

4. Requirements for a Digital Document Record System

4.1 Introduction

Many documents, within Higher Education, particularly those created by the administration, serve as records of the activities or decision making of the institution. In the past these documents were created either directly on paper with a pen, or, since the 1890s, with typewriters, and recently with word processing software. Many people question for how much longer paper can be considered the only medium of record, as the documents produced by computers lend themselves less and less well to printing out. For example, the print out of a spreadsheet does not show the calculations embedded in it. And web pages are frequently produced dynamically 'just in time' for the individual requesting the information, so it is quite possible that **two different people viewing them just minutes apart will see different information**.

Against this background, there is a pressing need to see the digital versions of the information as the record and not have to print it (de-digitise it) to create the enduring record. Moreover, in some jurisdictions it is only permissible to preserve one rendition of the information, either in the paper or in the digital order.

This section compares a number of approaches to the creation and use of digital documents as an enduring record of activity and decision.

4.2 Record Creation

4.2.1 Document Context

At the point of creation, metadata needs to be added to identify and define the document, providing its context, its purpose, where it is located and the (automatic) management of its retention and disposal.

The document creator is the person most likely to know the full history and ecology of a document. For this reason, if metadata is to be attached to a document, the creator is best placed to do this. The amount of research required and the manpower needed retrospectively to 'catalogue' documents at some time after creation makes it unlikely that resources would ever be found to do so. If tools (to make the addition of metadata at the time of creation a simple and time efficient process) are made available then there is a reasonable chance of producing documents with sufficient metadata to make their long-term management cost-effective and straight-forward.

Most documents are intended to impart information to people other than their creator. Increasingly, digital circulation is becoming important, although many people simply print off the digital copies that they receive. This requirement for digital circulation has resulted in two forms of practice within organisations:

- sending word-processor format files (e.g. *Microsoft Word* files) as attachments to e-mail messages
- transforming word-processor format documents to *HyperText Markup Language* (HTML) and placing them on a Web server.

In either of these ways of distributing digital documents, the standard mechanisms for doing this preserve little in the way of metadata, except the context within which it is sent (in the first case) or placed on the Web (in the second). The document is plucked from a directory file structure, which probably communicates a great deal of context in lieu of metadata, on the creator's machine. In the former case it arrives at a recipient's machine with a filename that probably conflicts with a name already in use such as 'report.doc' and may or may not be stored by the recipient in a location which recaptures some of its context (what, when, where, etc.). In the latter instance some context will be supplied by where the document sits in relation to other documents in the physical or logical structure of the Web server, but this does not always occur. To have value as a record, much of the metadata implied by its location in the directory

structure on the creator's machine or the web server must be captured explicitly. One important reason for this is simply that it cannot be guaranteed that this directory structure will survive the transformations of the documents necessary to out-pace technological change and ensure their long-term survival.

If metadata is to be added to documents then the only cost-effective time to add it is at the time of document creation and for the document creator to be the person who adds it. A digital record document creation system must have the facility to capture necessary metadata manually and preferably to automate the process as much as possible, for example by adding repeated information such as item numbers, location details and so on.

4.2.2 Document Structure

Most documents have inherent structure, which varies according to document type. As has been discussed earlier, in many documents produced with word processing software, much of that structure is coded only by changes in font size, weight, etc. This is not because the software is not capable of capturing this structural information, rather that the way in which many people have learned to use such software has not exposed them to these capabilities. Presentational cues to structure may not survive the transformations necessary for the long-term preservation of the records, for example the definition of heading types.

A digital record document creation system must have the facility to capture document structure manually, but preferably it should automate the process as much as possible.

4.2.3 Document Identifiers

To be useful as an information base, units of information in collections of documents require clear, consistent and unambiguous identifiers whatever their storage medium. These units may be entire documents, collections of documents, or component parts of documents.

The informal and unstructured forms of identifier and reference often used by authors ('the report of the standards committee', 'the minutes of the previous meeting') perform their function effectively only because the 'scope' of the reference - the range of documents from which the specific target is to be identified - has been limited. This may be by physically bringing together a subset of documents into a single 'file' or 'folder', and secondly, the human reader of the reference adds additional information drawn from the context of its use in order to determine its intended target.

A related problem arising from not assigning clear and unique identifiers (and names) to documents is that when they move from one person's file space to another, they undergo name changes and/or name clashes with existing documents, which results in confusion and inadvertent loss of documents.

If a computer program, be it a mechanism for processing of document records or simply a document viewer such as a web browser, is to be able to resolve a reference effectively, then a greater degree of precision is required.

Every document should be assigned an identifier by its creator at the time of document creation and a filename related to that identifier.

4.2.4 Integrity and Authenticity

If a document is to be considered as a record, users of that document must have confidence that the information contained within it is what it purports to be and has not been altered or corrupted at any time. This issue has both technological and procedural strands. In the paper world, the features that provide these guarantees are mostly procedural, involving where, how and by whom the paper record is stored, how it got there and who submitted it for storage. A holographic signature has traditionally been placed on the 'certified' record of proceedings of committees, but that is not necessarily a requirement for authentication of a paper record, it simply asserts that that individual was involved and this is only of value if the signature can be determined to be genuine. It is the statement of the policy surrounding the signing of the

document that tells us what significance we may attach to the signature and the fact that the document is signed.

There are considerable technological challenges associated with integrity and authenticity in the digital order. The procedures by which documents are created, stored, distributed and archived are what will ensure that the document serves as a credible record of activity and/or decision making. This requires more explicit procedures than have so far been available for the creation of digital documents. The role of digital signatures in the assertion that a document existed in some particular state at a particular time is not clear in the digital world where the document is likely to undergo transformations to out-pace technological change and to accommodate changing usage. Again it is the policy surrounding such signing that is important in establishing value.

A digital record document creation system must be set up in such a way that the procedures for document creation and management are such that the documents within it remain credible as records.

4.2.5 Adding Value with Digital Documents

In documents on which the ERM project has concentrated, making the structure explicit is a first step to the documents adding value to the processes of which they are part. A document in the digital order can and should be used for further processing within the transactions of which it forms a part, rather than simply being a step in the creation of a representation of information on paper. In the paper order documents do not have this potential. If advantage is not taken of these possibilities, we are simply using a more expensive set of tools to create the documents that we previously created with typewriters, with few advantages to show for the additional expense.

Providing adequate metadata facilitates the provision of discovery tools which can make use of the explicit information about the context of a document, its relationship to other documents and the relationships between bodies producing documents. Information within documents and about documents can be an essential part of the work-flow encompassing the movement of documents between individuals within the system. For instance, a proposal document often has to pass through a number of bodies who add value to it and pass it on to the next stage of development or approval. Such a document should carry with it enough explicit information for it to be automatically routed from stage to stage. Documents which contain actions should also have enough explicit information to allow a process operating on them to automatically remind at specific intervals the individuals on whom the actions rest, that they are still to be completed.

A digital record document creation system should add value to the process of which it is a part by making structure, context and relationships explicit.

4.3 Record Use

The most immediate use of documents is to communicate information between people or to record an activity such as the decisions made by a committee. In the committees, documents provide members with information about the issues to be discussed, background information about those issues and a record of what has been discussed and decided upon. In many cases documents will be used as a source of reference by those wishing to confirm the details of an activity, decision, etc. hence their longer-term value as records. In the past, most of this material has been stored on paper in filing cabinets, box files, etc. Latterly the word processor file which was used to create the paper version is often also in existence, but may not be considered in most people's minds to be the 'original'. This is because few people understand that the very act of saving a word processed file produces a permanent copy of the current rendition.

4.3.1 Paper Versions

Paper versions of record documents are usually in the hands (or filing cabinets or safes) of the major players involved in the activity, with perhaps surrogates available for reference in a library or 'originals' stored as an authentic record in an archive. The record has often not been

readily available to a wide audience, not because of any real need for secrecy, but quite the reverse because a circulation on paper to those closely involved was deemed too costly. Such convenient pleading of confidentiality has often been overlooked once other methods of distribution became available. It has been assumed that custom and practice dictates confidentiality, rather than a specific requirement. Freedom of Information legislation will require that reasons for confidentiality are examined rather more closely. Items which deal with specific individuals (and thus are covered by the Data Protection Acts) or are commercially sensitive are kept confidential (or 'Reserved').

In some circumstances many people involved with the creation and immediate handling of records find it necessary to have some material printed out on paper, so that they can read it on a train or by their fireside and have it to hand in a meeting. This is likely to be the situation, until either:

- there is a digital document reading device which is as easy to carry, annotate and use as paper, or
- laptop computers and connections to the network (either wired or wireless) are ubiquitous and their price and weight has fallen dramatically.

A consequence of this is that records systems will continue in a hybrid paper/digital condition for the foreseeable future and this has to become an accepted part of planning for the transition from paper to digital records management. Attempts to ignore this reality have resulted in the breaking of systems which have served for many years, whilst not delivering significant advance. This in turn has reinforced user resistance, leading to friction and in some cases rejection of the digital.

A digital record document creation system must take account of the fact that paper and digital representations of records will have to co-exist for the foreseeable future

4.3.2 Web Versions

Versions of documents that have been transformed to a form suitable for viewing on an Intranet, enable a wider spectrum of individuals to have access to them and to help both in informing and in dissemination of decisions. It is now almost *de rigueur* for many documents produced in Higher Education to be available on a Web server, either for restricted internal use or for wider dissemination. Even if some people have a paper copy initially, it makes sense for them to dispose of the paper copies once the activity is concluded. Then they simply refer to the documents on the Intranet as and when required, otherwise many different individuals stuff their filing cabinets with documents they might rarely need. This presupposes that there is some guarantee that the documents will remain available for an agreed length of time (perhaps in perpetuity). Decisions of this nature are commonplace to records managers, but less well understood by Webmasters. If the practice described above is to be established, there has to be clear policy in relation to maintenance and retention of records placed on Intranets (or other types of Web site). People can then destroy their personal copies, confident that they will be able to gain access to an authoritative copy as needed. Consequently it must be agreed that the *World Wide Web* (WWW) copy is 'authentic' or else a faithful representation of the original.

If digital records are to be made available via an Intranet, there needs to be a clear policy on the maintenance and retention of those Web versions, so that people can destroy their personal versions, safe in the knowledge that they will be able to refer to an authoritative and authentic copy if the need arises.

4.3.3 Other Digitised Forms

Digital documents always provide a mediated experience of the record, so it follows that it will always be necessary to have forms which are available to enable users to examine a record. It might be that a range of representations, including the original word processed form entered by the creator, are made available. The shelf-life of each of these representations depend on whether or not it is in a proprietary format, how close to the end of its life the format is and a range of other factors all of which are outwith the control of those responsible for maintenance of the record. People within organisations (particularly Higher Education) have different

software environments and it cannot, therefore, necessarily be assumed that one particular digital format will be suitable for everyone or indeed for every purpose. Some digital records are likely to need to exist in a number of different forms and defining suitable procedure as well as technological practice for making those transformations is urgently required, so that users of the documents can have confidence that the representation which they are looking at is not materially different from the representation created in, for example, a word processing package.

Decisions on the appropriate digital forms of record documents cannot be left to the vagaries of software supplier decisions on proprietary formats, there is likely to need to be transformation from format to format as technology changes. Some formats will survive longer than others and therefore have different intervals between transformation. Procedure as well as technological practice is again important for survival transformations, so that there is confidence that the record does not change materially over time.

A digital record document system must have well defined and understood procedures for transforming the record content from one representation to another.

4.3.4 Other Processing

Given that documents contain structure, the different elements of a document, such as a list of committee members present or absent, meeting items and sub-items, disposals, actions, tables of contents, etc., if explicitly identified within the document, could be used to drive processes such as automatic reminders to people with outstanding actions, automatic passing of matters to a higher level committee, tracking of item progress through a committee system from first initiation through to final approval and so on. Documents are, in some ways, a kind of loosely structured database with each piece of text falling into a specific type of 'field' and each 'field' having a set relationship to other 'field' and types of 'field'. In the text processing world these 'fields' are called elements and are used throughout the publishing world, as containers for pieces of content stored separately from their presentation for maximum flexibility.

If digital records are to achieve their potential to improve process as well as provide simple access to visual representations of content, then forms which allow such processing must be produced directly or indirectly from the initial document. In addition there should be no requirement for manual modification to enable the additional processing to work on the document, otherwise the benefits will not be fully realised. This implies that structure information must be explicit in the document and not only discernible by presentational cues. For example, to return to actions within minutes it has become common practice to list these in the right hand margin. In a digital representation it should be possible, as already stated, for these actions to be a dynamic list. Similarly embedded within a set of minutes will be references to the same topic made evident by the heading again it should be possible in the digital world to treat these dynamically.

Structural elements in a digital record document should be explicit so that further processing may be added to such systems without manual re-working.

4.3.5 Access

Control of access to documents or their constituent parts, such as confidential or otherwise 'classified' documents or the reserved business sections of committee minutes, is a major issue. It has important implications in relation to the Data Protection Act, security of the institution, commercial sensitivity and so forth. A requirement of any system which makes records available is that an appropriate security model is specified and implemented. A problem which is frequently encountered, even when such a security model has been defined and implemented, is that the control measures are generally not under the control of the document creators. They may require the technical assistance of a systems' administrator to deal with access control and, in many cases, also to deal with the publication of the documents on the Intranet. Until institutions deal effectively with these problems, document creators will not be in a position to comply with the security model and are either unable to make documents available when they should be or will risk making a confidential document visible to a wider audience than they should. In some ways the digitisation of records makes control of access more, rather than less, complex for document creators.

A digital record document system must be underpinned by a security model and an effective implementation of that model, which allows document creators to share documents without having to negotiate unnecessary hurdles which lead to circumvention of the security model.

4.4 Record Disposal/Preservation

Part of the metadata captured at document creation should indicate the retention schedule for the document. For many of the decision making committees in universities, the records of the decisions might be expected to have a relatively long life, but, whatever the appropriate period of retention, it needs to be explicit in the metadata so that automated processes can deal with the transfer from inactive record to archive when the time comes. For some documents of record, there will be statutory limitations governing the time the information should remain available and for others there will be requirements imposed by the institution or by external agencies that after a particular length of time the document must be destroyed. Somewhere this retention information must be made explicit if automated processes can take over some of the work involved in ensuring compliance.

4.4.1 The Archival Process

The digital order must seek to replicate the physical activity of 'archiving' in the paper order - in other words the transfer of the 'control' of the asset (or at least an authentic surrogate) from the creator into independent guardianship. Because of the permanent possibility of alteration or degradation of the original, transfer will need to take place as soon as the transaction has been completed. As in the paper order, the transfer will need to be accompanied by any associated documentation. Such transfers will need to be documented more thoroughly than at present to prevent subsequent arguments about authenticity.

A digital record document system must undertake immediate, auditable transfer of a digital record, designated as archival, on creation into the digital archive system.

4.4.2 Verification of Records

In the paper order a minute is normally verified by the signature of the chairman following a minute at a subsequent meeting that the contents were a correct record of what took place and if they were not, what amendments were made. This process will need to be replicated in the digital order. Since corrections are more difficult to make explicit in a digital form, practice may have to change to include a minute of changes.

A digital record document system must have procedures in place and should have the facility to verify and record amendments to documents and possibly data elements.

4.4.3 Access to Records

Rights of access to information depend both on the nature of the record and the content. Some transactions are a matter of public record and others are not. The boundary between the two areas is shifting with the growth in consumer rights and accompanying concerns about accountability. Nevertheless some records will, of necessity, have to remain confidential, for example any transaction relating to an individual for at least the person's lifetime. The interface between the Data Protection Act 1998 and the Freedom of Information Act 2000 typifies this difficulty. It is not difficult to manage such 'rights' in the paper order but it is much more complicated in the more open digital order. It is essential therefore that the metadata includes information about such rights along with a description of content, as it is likely the legal requirements to ensure 'rights of access' will change.

A digital record document system must have an effectively implemented access model which safeguards confidentiality and privacy rights.

4.4.4 Verifiable and Tracked Disposal

In an increasingly litigious world, institutions need to be able to demonstrate with certainty that records, which are time expired, have been destroyed. This is not easy even in the paper order where record keeping is devolved unless there is a strong compliance culture, which is not the case in HEIs. In the digital order it is even more problematic as many companies and organisations have discovered. This is because many renditions of the same document are likely to be preserved (for example on backup tapes) even if the 'original' has been destroyed in good faith.

A digital record document system must have procedures to ensure that all renditions of a document are destroyed when they are time expired and be able to document the process.

4.5 Long-Term Survival of the Physical Record

The media on which digital material is held have a distressingly short life-expectancy. Virtually everyone will have had the experience of a floppy disk that worked one year failing to deliver the files which are on it the next. The life-span on CDs is not well understood, but for the type which can be written by end-users, it is thought to be at best a few years. In addition, file servers are replaced for time to time and there is a need to move files to new systems to cope.

The net result is that there is a need to move digital files to new media on a regular (if not frequent) basis. Checks on the integrity of files, which verify that the bits which make up the file have not changed, will be required both as a general check on the 'health' of the storage medium, but also that the medium transfer has been accomplished with no degradation of the file and therefore the integrity of the information contained. Moreover these inspections must be carried out by trusted, independent third parties who have no interest or nothing to gain by altering or destroying the content.

A digital record document system should have integrity checks on the digital objects as an ongoing 'health' check, but also before and after medium transfer.

5. Glasgow Committee Process and Systems

5.1 Introduction

The following chapter sketches the nature of the environment at the University of Glasgow, and the function of the committee system within the organisation. It seeks to highlight elements of the technical, organisational and 'cultural' context that conditioned the approaches taken by the project, and to narrow the focus to the particular requirements of the committee system and its use of digital records.

The project chose a subset of the University's committees as a test-bed. All documents submitted to the meetings of those committees are designated as records, that is, they served as evidence of the activities of those committees. They allow people to see what was decided, when, by whom and, to an extent, how they arrived at the decision and who was consulted. The project undertook to develop tools and protocols to facilitate the management of these records in digital form, from the creation of the documents, through their distribution and use, to their final disposal.

The principles developed by their investigation have a broader applicability across the university and for the sector.

5.2. Committee papers at the University of Glasgow

5.2.1 The University of Glasgow and its committees

Like many similar institutions, the University of Glasgow has a large number of committees. Some, like the Senate and the University Court are both high profile and influential. The remits of the Court and Senate themselves are clearly defined in the Universities (Scotland) Acts 1858 and 1889 - broadly summarised, the Court is responsible for the resources of the University and acts as a court of appeal against the decisions of the Senate, and the Senate is responsible for the superintendence of teaching and research. Their committees maintain this distinction, with Court Committees concerning themselves with resource matters, such as the estate, staffing matters and finance, and with Senate Committees considering educational policy, research matters and student discipline.

However, it could be argued that the differentiation is no longer as valid as it once was. The Jarratt committee[15], which examined the efficiency of the university sector, recommended bringing together academic and financial decision-making as near the 'chalk-face' as possible. This was interpreted in different ways. On the whole the new structures that were put in place simply overlaid the traditional hierarchy of committees. As a consequence, in the University of Glasgow there are three 'streams' of committees involved in decision-making - Court and its committees, Senate and its committees including Faculties, and the Resources Strategy Group composed of deans of faculties and the Management Group. The executive management of the University's business is largely conducted through this third stream, by-passing the other two streams which handle formal business. Relations are uneasy and the balance of power is continuously shifting.

Prior to a review in 1998/9, there were approximately 68 committees of Court and/or Senate (some being joint committees of Court and Senate).

5.2.2 The Information Committees

The information committees consist of the Information Services Committee, which is a joint committee of Court and Senate, and a series of sub-committees of ISC including:

- The Library Committee

[15] *Report of the Steering Committee for Efficiency Studies in Universities*, Committee of Vice Chancellors and Principals, 1985.

- The Information User's Committee (IUC)
- The Advisory Committee on Standards, Guidelines and Protocols (now subsumed into IUC)
- The Information Strategy Steering Group (prior to the 1998/9 review)

5.2.3 The information landscape

The work of the ERM project was directly conditioned by three prior or existing initiatives within the University:

- the University of Glasgow's Information Strategy
- the experience of paper-based Records Management practice
- existing electronic document management initiatives

The University's Information Strategy, published in mid-1997, emphasised the need to make more **effective** use of the university's information resources. It highlighted the fact that the information flow both within the university's committees and outward from these committees to other members of the university was an area in which digital dissemination techniques could offer scope for greater efficiency, in terms both of committee members' time and the time of others wishing to reference the decisions of those committees subsequently.

In mid-1997, the *Records Management Team* (RMT) based in the University Archives began a survey of the University to draw up a retention schedule with the objective of classifying records so that unnecessary records or those that no longer need to be retained are destroyed in a systematic manner at the end of a specified period and those of continued value are given treatment appropriate to guarantee their preservation. The RMT operates a Records Centre to provide secure storage and efficient retrieval for those paper records which are of continuing value but which their owners/creators do not need to store locally.

A number of other projects, in various parts of the University, have been active over the past few years. Principal amongst these is the Senate Documentation Project and its predecessor the Science Faculty Document Project. The ERM project has learned much from these other projects, which have had many broadly similar aims, but have not developed the records management context to the same extent. The Senate Documentation Project has now been subsumed by the successor to the ERM project.

5.3 The flow of information in the committee process

The flow of information in committees is based on a cycle of document production and circulation starting with an agenda and finishing with a set of approved minutes.

The first step is agenda production and circulation of this to committee members. There may be a number of iterations before the agenda is finalised and the items for discussion identified.

Agendas usually reference supporting papers that have been submitted for consideration by the committee in its deliberations. Some papers originate from individual committee members; some come from subordinate committees, from departments, or from other individuals within the University; and others originate from bodies external to the University and sometimes bear their own schemes of reference particular to the creator. The production (for those authored within the University) and circulation of these papers is the second element in preparation for a committee meeting.

The documents provide the members of the committee with information about the issues to be discussed, background information about those issues and a record of what has been discussed and decided upon in the past. The majority of committee members bring printed copies of these documents to the meetings - if not all the documents for the meeting, then certainly a subset.

After the meeting has taken place, the clerk produces a set of minutes based on the structure of the meeting identified in the agenda. This document is then circulated in draft form for the members of the committee to comment on and suggest changes, corrections, etc. At some time

later, **either** at some point between meetings when members have had a period to reflect on the draft, **or** at the next meeting of the committee, the minutes are adopted as a true and accurate record of the meeting and may be signed by the committee convenor or chairperson.

It is common for staff within the University - including non-members of the committee - to continue to make some use of these documents after the business of the individual meeting has been concluded and its minutes have been approved. References to minutes and papers may be, also, used to publicise the decisions of the committee; they will certainly be used as a source of reference by those wishing to confirm what was decided by a particular committee at a particular time. There are cases, however, where access to such documents may be restricted to members of the committee.

Both minutes and papers may be referred to in later meetings of the same committee, or form the basis of submissions to other committees at a later date. For example minutes of many committees are tabled at meetings of Court and Senate where questions of substance may be raised and therefore minuted.

A paper which was submitted to a committee may subsequently be drawn to the attention of a different user constituency. In some cases, this sort of 're-use' may result in the authoring of a separate document for that purpose - perhaps with only minor modifications to content, but distinct nonetheless; in other cases, a reference to the original may suffice. Changes to courses are determined by Boards of Studies for later ratification by the appropriate faculties or approval by Senate. In the paper order this produces at least three renditions of the original decision.

In most cases, the availability of committee documents serves to validate the legitimacy and accountability of the decision-making process itself. After a period of time has elapsed (usually when the filing cabinet is full), it is recognised that the records are not 'in active use', and the records are disposed of - either destroyed or transferred to the archives for preservation as a permanent record.

5.4 The role of the records manager

The records manager provided a training resource for the project to ensure that record creators had some understanding of records management principles across the range of their work and between the paper and digital domains.

The remit of the team was threefold:

- review current records and information management procedures across the University, recommend improvements and design appropriate strategies and systems
- establish an integrated information management system for the Principal's Office
- devise an externally marketed training course in records and information management for those, without formal qualifications, working in the field

The construction of a generic retention guide so the volume of duplicated records in the university was reduced was a key requirement for managing the custody and longevity of digital records.

5.5 The creation and distribution of the Information Committee's papers

Prior to the initiation of the ERM project this flow of information was accomplished in the following ways.

The committee records were created primarily using desktop word-processor tools, and staff from a wide range of academic-related, clerical and secretarial grades carried out the work. The emphasis was still on the production of presentable paper documents. Some high-level guidelines existed for the presentation, structure and content of committee documents; in practice, however, they were not widely implemented or adhered to.

Although the devolved and plural nature of the institution means that the precise nature of the individual record creator's environment varies, the range of tools used is quite limited. Opportunities existed to shape the use of those tools, both through guidance and training in

'good practice' and through the provision of specific software routines to perform specific functions.

Records were distributed in a variety of media and representational forms:

- hardcopy printed versions (either circulated in advance by internal mail or 'tabled' at the meeting)
- digital representations encoded in proprietary word-processed formats (principally *Microsoft Word* documents circulated as email attachments)
- HTML-encoded forms distributed via the university's Intranet

This latter class of representations was, in the case of the Information Committees, hosted on a single server maintained by the University Computing Service. The creation of the HTML-encoded versions was carried out either by the author of the document content or more often by another member of staff who had some knowledge of HTML, and usually involved some 'cut-and-paste' of the content of a *Microsoft Word* 'source' document.

The documents were stored by the committee clerks in the short-term both in hardcopy and in proprietary word-processed format. Although committee clerks were encouraged to be consistent and methodical about the physical storage of digital documents within a (sometimes shared) hierarchical filestore, little had been done to 'manage' the digital forms of these records and only the printed versions were considered suitable for archival purposes.

Although both document creators and document users recognised the potential for improved means of distributing and preserving the documents of the committees, for the most part, creators continued to employ approaches most suitable for text destined to be printed and used in hardcopy form. Indeed, this was simply an accurate reflection of the use of those documents: it would be fair to say that even where they were **distributed** digitally, the documents were **used** first and foremost in printed form. Many committee members either requested the distribution of paper copies or printed off the digital copies which they received in their entirety. All this achieved was the transfer of cost from the secretary's photocopying budget to the more expensive process of local printing. This was sometimes the implicit reason for adopting electronic distribution, particularly for very large committees such as Senate.

5.6 The CDocS tools for the creation and distribution of committee papers

The challenge faced by the project was to:

- encourage and enable the document creators to adopt practices which
  - respect the principles of good records management practice by being internally consistent with referencing systems independent of the representational form, and supporting metadata which describes the precise location of the instance of the document in the overall structure of record keeping,
  - generate digital representations which are durable and reusable,
  - supply the additional data that facilitates their long-term management and provides an intellectual and administrative context sufficient to give them value as records
- provide a framework for the use and distribution of these records which
  - ensures they are delivered to the intended recipients by guaranteeing that membership/circulation lists are current,
  - meets existing requirements for the functioning of the committees as decision-making bodies,
  - addresses issues of availability and access control as demanded by the nature of the information and the practices of the committees and any external constraints,
  - contributes to more efficient functioning of the committees by providing a flexible digital information base,

- provides for the appropriate disposal - destruction or retention - of records at the end of their period of active use however determined,
- provides for verification of the integrity and authenticity of the records,
- dovetails with the practices established by the *Records Management Team* for records held in paper form,
- establishes the requirements for a program of long-term preservation for these (and other) records and to ensure that any procedures implemented were in line with those requirements.

It was recognised that with appropriate training the document creators are in a position to supply much of the information required in such a system. Committee clerks in particular play a crucial role, as creators, distributors and, in many cases, 'custodians' of the records of the committee. For this reason, the project has concentrated considerable efforts on shaping the practices of record creators, and has supplied a set of document creation tools (a collection of *Word* templates and macros) which provide a semi-controlled environment for the document creator. These tools are referred to as the *Committee Document System* (CDocS).

The aim is to facilitate:

- a structured approach to document creation,
- the capture of basic metadata values,
- some standardisation of document content,
- standardisation of presentation,
- the distribution of documents in a manner consistent with the requirements for access in a compliant environment.

5.6.1 Document creation

The project has created two sets of document creation tools for use within the *Microsoft Word* environment. They consist of a number of templates which provide structural outlines for the basic document types and a supporting set of macros which control the use of those templates and provide dialogues for the capture of basic contextual information which is associated with the document.

The first set of tools manages the creation of agenda and minutes documents and is intended for use by the committee clerk; the second is designed for the use of those submitting papers for consideration by the committee.

5.6.1.1 Agenda-minutes wizard

The process that directs the creation of the agenda and minutes documents is based on the proposition that both these documents have the same outline structural form. They may, however, be formatted quite differently, and within that outline structure, the minutes document for a given meeting will contain significantly more narrative text content than did the agenda document for the same meeting.

The agenda-minutes wizard presents a series of dialogue boxes to collect basic information about the committee meeting - the name of the committee, date, time and location of the meeting, and the list of items to be discussed.[16] The values for some of these properties can be selected from pre-defined lists. The wizard then uses that information to create a skeleton document which is presented to the clerk for further editing.

At the start of a committee meeting 'cycle', the clerk executes the wizard 'in agenda mode', supplying the meeting details and the headings of items for discussion, and generates a skeleton agenda document. After the meeting, the wizard is re-run to retrieve the agenda data (including items added by the clerk during their editing of the agenda document) and a skeleton minutes

[16] See Appendix 1 for some examples of the agenda creation dialogue.

document is created.[17] This leaves the clerk the job of adding the narrative account of what has been discussed, the decisions made and the actions agreed.

In the cases of both the minutes and agenda documents, the meeting information supplied during these dialogues is retained with the documents as the values of metadata properties. The macros also add some further properties that are required for the management of the record.

5.6.1.2 Paper wizard

A similar tool is available for the creation of papers to be submitted to a committee meeting.

A dialogue requests basic information about the meeting to which the paper is to be submitted, and that data is retained as the part of the document's metadata.

5.6.1.3 Editing the skeleton document

The approach taken by the ERM project depends on the creator employing the *Word* 'styles' made available from the document template in a carefully controlled manner, not simply to apply presentation consistently, but to describe the logical structure of the document.[18] The wizard process sets display and edit options within the *Word* environment to make this process as easy as possible, and additional support for the author is provided by macros to perform common functions, accessed via toolbars.

Nevertheless, such an approach does require a shift in practice, particularly where document creators have been accustomed to placing an emphasis almost exclusively on controlling the layout of documents on the printed page with little reference to any shared standards, even for presentation. Even for authors who are familiar with the use of *Word* styles and templates, the approach necessitates a rigour in their application which may be unfamiliar. In recognition of this, the project has directed considerable efforts to encouraging users of the tools to think 'more structurally' about the process of document creation, and to provide detailed supporting documentation for the tools.

The complexity of the structural models which can be accommodated by this technique is limited. In some cases, this has required authors to simplify their practices making them consistent with the constraints of the tools. In the majority of cases this simplification has resulted in better structured documents of greater clarity.

Furthermore, the approach requires that all those authors creating documents which are submitted to the committees implement the CDocS tools, and can 'fit' their documents within the framework of the structural models provided. In fact, the number of individuals who submit papers to the Information Committees is relatively small, and it has been possible to encourage a high proportion of paper creators to adopt this approach. In practice, however, there are always some documents which either originate from external sources or have structures which do not fit within the CDocS models. In these cases there has been little alternative but to adopt an approach which satisfies the short-term requirement for distribution by generating *HyperText Markup Language* (HTML) -encoded representations which reproduce more or less the presentational characteristics of the word processed forms.[19]

5.6.2 Document distribution

Documents created by the process described above are distributed by uploading to the Intranet web server (hosted by the University's Computing Service), where they are subject to a process of validation and transformation[20] which generates a number of representational forms. Committee members and other users access these outputs via a Web browser program.

---

[17] See Appendix 1
[18] The approach to document creation is discussed further in section 6.4.1.
[19] This is characterised as the 'presentational approach' in section 6.2.
[20] See Appendix 2 for more details of this process.

The transfer of committee papers from local storage to the Intranet server is controlled by the committee clerk. Transfer is initiated through the use of an HTML form which allows the clerk to select the (locally saved) *Rich Text Format* (RTF) representation of the record. As a rudimentary safeguard against the accidental selection of an incorrect document, some other basic descriptive properties are requested at the time of upload, and the transfer process validates the user-supplied values against the values of the corresponding metadata properties embedded in the selected document. Access to the transfer form is password protected.

If a document goes through several versions, as is frequently the case for both agenda and minutes documents, the clerk simply re-runs the upload process to replace the previous versions of the representational forms on the server. Considered as a record, it is not necessary to retain the various draft forms of the document: the only version which is of value as evidence of the committee's activity is the final version.

The clerk notifies committee members of the availability of documents by email.

On the server, the documents created in the CDocS system are converted from their RTF representation into an Extensible Markup Language (XML) forms for both the document itself and its associated metadata. The XML forms are available for archiving and are also used to generate the representations which the committee members see.

Currently committee members are offered an HTML representation both of the document itself and its metadata directly on the Intranet web server. In addition the RTF form of the document is made available. It is possible to create other representations directly from the XML representation if there is sufficient demand to make this worthwhile. For example we have done work to generate Portable Document Format (PDF) representations, but currently do not offer these to committee members. Note that these PDF representations are generated from the XML and not generated directly in the word processing environment.

In addition to the representations of the 'whole' document, already discussed, there is considerable scope for representations of only part of the information. An example of this might be an abbreviated form of the minutes which simply contains the item headings and the disposals and actions. The committee convenor or clerk could then use this when setting up the next meeting or reviewing progress.

Documents not created in or converted to the CDocS system are simply converted to HTML or uploaded as they are, but a basic set of metadata is collected manually as part of the upload process.

5.6.3 Document storage and access

The process of transferring a representation to the Intranet server via this controlled mechanism takes the record from the clerk's local document creation context and places it within a managed environment for active records. The representations of the record which are generated as part of this process are recognised as the 'authoritative' versions and are made available via the web server.

The significant factor is that the records and their representations are brought under the control of a 'management system'. This should not be confused with the fact that this step happens to coincide with their being made accessible via an *HyperText Transfer Protocol* (HTTP) server accessible to the user constituency. Indeed bringing the record under control of such a system and providing access to it are two quite **separate** functions, and it might be argued that it would be helpful for the ERM implementation to establish a clearer distinction between them. It may well be the case that the clerk's immediate priority is to 'put the minutes on the Web', but from the records manager's perspective, establishing early control of the active record is equally important. 'The Web' is simply one means amongst many of making representations of the record available to a group of users. (It is also worth noting that the same physical server which is used to store these committee records no doubt contains both (a) digital objects which are not records and (b) digital objects which should be considered as records but have not been brought under systematic control in the same way.)

Committee members and others download copies as surrogates from the Intranet server for use. The life-span of that downloaded copy may be as brief as the time taken to read it once on the screen display of a web browser, but in the majority of cases this tends to mean generating a

printed copy of one of the representations available. This is likely to remain the case until there is a digital document reading device which is as easy to carry, annotate and use as paper.

Although this outcome is still well in the future a few committee members make use of the downloaded representations in their digital representation, perhaps by transferring them to a portable device which they take with them to the meeting in order to view the documents. There are risks associated with such practices as loss or theft of such devices can seriously compromise security and leave the 'owner' open to serious allegations just as the loss of a briefcase full of papers would do.

Whether these copies are in hardcopy or digital form, it is quite likely that their users retain them for some period of time after the meeting - quite possibly for much longer than they are actually making use of them.

If it can be guaranteed that the representations available from the server will continue to be available for an agreed length of time - and depending on that period of time, such a guarantee may require an undertaking to generate new representational forms as determined by changing use environments - then users can destroy their locally-held copies (in whatever medium) with the confidence that they will have continued access to the stored record. Given the liability contingent on the holding of information locally, this must be an objective of any system for the electronic distribution of documents. Therefore in order to ensure privacy concerns are met it is necessary to ensure that a secure deletion process is in place.

5.7 Conclusions

The committee process is capable of absorbing almost any amount of resources without limit. Considerable expenditure on IT equipment and software is made every year by universities. It is not clear that working practices are changing to take advantage of the opportunities that the new technologies bring. If this does not happen, then institutions are in danger of using expensive technologies to carry out tasks which could be carried out to the same standard and with the same 'limited' benefits as could be achieved with less sophisticated technology.

A major driver for change is that expectations of committee members are clearly changing; they are starting to expect to be able to access material on the web and not to have to keep filing cabinets of papers to which they will refer infrequently, if ever. They are also expecting to be able to transact small amounts of business without face to face meeting. If transactions take place by e-mail, it is important that there are systems in place to ensure that all members are involved in the transactions and that an appropriate record of the transactions is made.

While the approach taken with CDocS can not offer the flexibility and control of a structured authoring tool, for simple structural models, techniques are available within *What-You-See-Is-What-You-Get* (WYSIWYG) tools to capture some description of the logical structure of the document, as well as formatting. That description of logical structure can be extracted and applied to the content in the form of an *Extensible Markup Language* (XML) document. In addition the emphasis on giving the committee clerks control of the entire process right through to publishing and archiving reduces the need for technical staff to support parts of the process. It speeds up the process and gives the committee clerks a much better level of task ownership and job satisfaction.

In CDocS, the approach of shaping the practices of record creators is critical. The benefit for the records manager - and for the organisation - is that digital records and their associated contextual data are brought within the control of a 'management system' at the earliest opportunity. In addition they are stored in representations which maximise their durability, reusability and disposal. These are not the only issues to address in the area of digital records management and preservation, but they provide a firm foundation on which to build.

6. Creation Strategies for Digital Documents

6.1 Introduction

Document creators have generally adapted the demands of printed form to the digital domain by opting to use *What-You-See-Is-What-You-Get* (WYSIWYG) authoring tools and techniques which give them the ability to predict and control the appearance of that printed rendition with no consideration of the underlying process. This was discussed in some detail in section 5.6. For the subset of these documents that are records of an organisation, it is the printed rendition which has been considered to 'be the record' and which has been treated and 'managed' accordingly. In this scenario, the digital representation of the document is of limited and transient value - in some cases, it may be used only by the author. It is simply a means to the generation of the printed form.

This chapter surveys some of the possible approaches to the creation of digital documents, and attempts to assess the extent to which those approaches satisfy the requirements for the documents which are generated to have value as records.

It should be emphasised that in the context of this discussion, the term 'creation' designates a process which extends somewhat beyond the generation of a single representational form as the output from an authoring tool - the proprietary form output by a word processor, for example. Here, 'creation' must encompass any subsequent processing required to generate the other representational forms in which the record is used during its active life.

In chapter 2 it was argued that several criteria must be met if a digital representation is to have enduring value as a record. The tools and techniques applied to the creation of digital documents destined for paper rendition must be examined critically in the light of these requirements. Of course, any demand to change or extend current practices must be balanced against the recognition that in most organisations document creators are familiar with a small number of tools and in their use of those tools employ a small set of (perhaps long -established) practices and techniques: any such change carries with it costs in terms of training. There may even be shifts in roles and responsibilities which carry broader implications for the 'working culture' of the organisation.

The discussion below categorises approaches to document creation into three broad classes. However, it is recognised that these distinctions are problematic: the boundaries are not always clearly defined and in some cases the techniques discussed exhibit characteristics of more than one of the classes.

6.2 The 'Presentational' Approach

This approach is characterised by an emphasis on the document creator's ability to control layout and presentation. The expectation is that those presentational qualities of the record will be preserved when the initial digital representation is transformed into other representational forms for use.

This emphasis dovetails with the expectations of authors accustomed to creating documents that are printed and used primarily in printed form, and tends to assume the use of WYSIWYG authoring tools such as word processors.

6.2.1 Creating Presentational Effects

Because such tools emphasise their capacity to control the layout of printed renditions, the range of functionality typically available tends to include many different ways of generating formatting effects.

i. Most tools allow formatting to be applied to text 'as required' at either the 'block' (paragraph) or 'inline text' (character string) level. That is, the author selects individual units of text and applies combinations of formatting characteristics (typeface, font size, appearance, indentation etc.) to those separate units as required.

ii. Even when the document **is** destined for immediate printing, this is not an efficient way of applying sets of formatting properties which are repeated throughout a document, and it tends to lead to irritating inconsistencies in presentation. Most word processor tools incorporate a feature that allows the author to store a collection of formatting properties used repeatedly in a document as a named unit so that they can be reused. In *Microsoft Word* this is the 'style' feature. Furthermore, collections of 'styles' can be stored outside the document (in 'templates', in the case of *Word*), so that they can be reused across a set of documents of the same class (or 'type') to facilitate consistency of formatting across multiple instances of that type of document.

iii. A third commonly used device for controlling the layout of text blocks on the printed page is the use of tables that can contain alphanumeric as well as numeric values.

6.2.2 The transformation of presentation

This approach relies on processes of transformation which take as input one representational form, probably held in a proprietary encoding format (i.e. containing procedural markup specific to a single rendition program which instructs that program to render text in certain ways), and generate a second representational form. The latter uses a different set of procedural markup conventions to instruct a different program to render the same document content in the same (or similar) ways. For example by importing a table from *Microsoft Excel* into a *Word* document where the data can be seen to be rendered in a different way.

Many word processing programs already contain (or can accommodate third party 'plug-in' versions of) 'filters' which perform this function. They allow the user of the program to read-in documents created using a 'foreign' encoding format and/or to write out in foreign formats versions of documents which were initially created in the program's 'native' format. Such filters may be designed to permit interoperability between different versions of the same word processing program or between the proprietary encoding formats of two different vendors.

Such transformation processes between two proprietary representation formats often reproduce the presentational features of the input imperfectly in the output representation. The reasons for such variations are rooted in the fact that the program-specific procedural markup languages almost inevitably incorporate techniques that make it difficult to establish simple translations between them.

Furthermore the use of presentational conventions is conditioned by characteristics of the target medium (see section 2.3.3.2). When the output representation is designed for rendition on a different device type or medium, the goal of generating direct translations of presentation may result in presentational features which, while more or less faithful to that of the input, seem slightly incongruous in the new medium. It may be necessary for the filter program to accommodate presentational features which have no direct counterpart in the output, as, for example, in the case of the rendering of page headers or footers in a representation designed for a single scrollable window. As long as the filter is equipped (or can be configured) to handle such contingencies, however, they do not provide an insurmountable problem.

As the Web emerged as a channel for document distribution, the demand increased for the ability to generate *HyperText Markup Language* (HTML)-encoded representations without manually inserting the HTML markup, and many word processors now incorporate an HTML output filter accessed via a '*Save as HTML*' function. There are a large number of third party conversion tools that perform a similar function.

Such transformations from proprietary word processed format to HTML, however, are notorious for generating documents which make use of HTML element types to achieve presentational effect, in a manner which is at best in conflict with the semantics of the HTML language and at worst is syntactically invalid. They often contravene the design goals of the HTML language in tailoring their output to the behaviour of a small subset of the browser programs in use - perhaps even the programs of a single vendor.[21] It is possible to devise 'work-around' strategies to accommodate such shortcomings, in the form either of pre-processing of the input document to ensure that it employs only those presentational features which are reliably processed, or of post-processing of the output to correct the errors which have been introduced.

[21] The World Wide Web Consortium HTML Home Page at http://www.w3.org/MarkUp/

The use of tables is a widely used device to attempt to gain control of the layout of HTML-encoded documents. Since filters preserve the structural and presentational characteristics of a table quite well in transforming between proprietary word processed and HTML-encoded representation, the tactic of 'pre-processing' the input representation so that its layout is described in terms of tabular structures can be quite effective in preserving layout.

Assuming that the complete transformation process (i.e. the execution of the filter plus any additional pre- or post- processing required) produces an output representation that can be correctly rendered, then it may be argued that this approach is satisfactory, **if** the (more or less) direct reproduction of the presentational features of the input representation is what is required - and it is **all** that is required.

Increasingly, however, the user requirements for a digital document transcend this, and the limitations of an approach based on presentation become apparent.

6.2.3 The limitations of processing presentation

The underlying limitation is that the only information (apart from the document content) which is available to the transformation process in the input document is the markup that describes presentation. The process does not have access to description of the logical structure which that presentation is intended to convey to the human reader.

Even some simple processing tasks require some manipulation of the logical structure of the document. The construction of navigational tools such as tables of contents relies on the capacity to distinguish structural components such as headings and pagination from the text that surrounds them.

In simple cases, it may be possible to establish an unambiguous set of correspondences between presentational characteristics and logical element types. For every piece of text which has been centred and rendered in font *Arial* and size 18 is a main title; every piece of text in font *Arial*, size 14 and face bold is a section heading; and so on. If these form-structure correspondences can be made explicit to the transformation program -for example as a secondary input - then the program can identify and process those components - to construct, say, a table of contents based on the identification of text formatted as headings.

But in practice, there is what Steven DeRose describes as 'an inherent asymmetry' between form and structure.[22] For example, italic face text might be used within paragraphs for several different purposes: to mark the first occurrence of terms introducing central concepts; to signal text as a non-English-language expression; and to highlight the titles of publications (books and journals) in the bibliography. While the human reader makes use of subtle and complex contextual information to differentiate between these uses, and so to divine structure from form, it would be almost impossible for a program to do so. Document form is generally determined by structure. However, while textual components which look the same may be structurally identically, this is not necessarily so: **structure can not be determined automatically from form**.

As a consequence, any more sophisticated **processing** of the document becomes problematic when the representation does not contain this structural description. It would not be possible, for example, to extract occurrences of subject terms or personal names as the basis of subject or name indexes, or to select those documents by a certain author or containing a specified expiry date value.

Increasingly, digital documents must be **reused** for purposes and in hardware/software environments other than those for which they were created. This often requires that they are transformed into a different representational form, and it was noted above that there are many filters and conversion programs which perform this function on the basis of reproducing (more or less) the presentational features of the input representation.

However, the range of user platforms, and the representational forms required, is permanently increasing. What is more, the character of those use environments is becoming more diverse, in

[22] DeRose, Steven J., 'Structured Information: Navigation, Access, Control', in *Proceedings of the Berkeley Finding Aid Project Conference. April 4-6, 1995*. (Berkeley Finding Aid Project, 1995). Available at http://sunsite.berkeley.edu/FindingAids/EAD/derose.html

a manner which means that the presentational conventions (which might include aspects of both formatting and physical structure) appropriate in one context are radically different from those in another. The growing use of hand-held display devices is a case in point. The physical constraints of such device mean that the presentation of a document requires the application of quite radically different presentational properties from those which might be used on a typical *Personal Computer* (PC) screen. It may be possible to establish one-to-one correspondences between the presentational properties in the input representation and those required in the output, but as the characteristics of devices diverge, this becomes more difficult to do reliably.

The real requirement for reuse, or 're-presentation', is to allow the logical structure of the document to drive the creation of the representational forms, so that the transformations operate on the basis of selecting logical components of the input and applying to them the formatting properties required **for the particular use context**. As has already been suggested, the direct replication of presentation designed specifically for one use context - typically that of print - may simply be quite inappropriate for that new context. And while it is possible to transform one set of presentational properties into another the 'asymmetrical relationship' between structure and form means that that may not be sufficient to address this problem.

6.2.4 The presentational approach and the digital record

Records are used for a variety of purposes. The full potential of digital records has yet to be realised, but they will undoubtedly be used in ways other than simply printing out an exact facsimile of the created input in the word processing program. If this is the case, a number of considerations apply:

- the records will be manipulated by automatic processes as well as by human readers,
- the presentational cues will be insufficient, by themselves, to carry the full structural meaning of the different parts of the record,
- records created using the presentational approach will not lend themselves to automatic processing because there is no simple one-to-one correspondence between the presentational features and structural components,
- considerable human intervention will be required to accomplish many processing tasks, as it is only the human reader who will be able to translate the presentation into logical structure.

The presentational approach is attractive, in-so-much as it is very close to the work pattern that most people have developed in moving from manuscript and typewriter document preparation to working with digital computer programs. Its limitations are starting to be felt in many areas of administrative work and the impetus for change will grow rapidly in the near future.

The following two approaches to the management of the digital record represent attempts to transcend some of these limitations by giving a higher priority - to a greater or lesser degree - to the explicit description of structure.

6.3. The database approach

This approach is predicated on a creation process in which the author explicitly divides the content of a document into component parts, and stores these parts in cells or fields so that those parts can, in theory at least, be accessed independently. Such a function can be performed using a number of different software tools, not just in database packages *per se*. It is clear that such a technique implies an approach to document creation that departs radically from the use of a word processing package, and training and culture change are implicit.

The extent to which the use of a relational database transcends the limitations of the strategy described in the previous section depends entirely on the 'granularity' (refinement) of the structure applied to the document stored within the database. As with some much else the right level of granularity is a trade-off between cost and benefit.

6.3.1 Representation stored as a Binary Large OBject (BLOB)

At one end of this continuum, one or more representations (perhaps encoded in proprietary formats) are stored as individual BLOBs, the internal structure of which are not further differentiated, and these BLOBs are associated with structured sets of metadata properties describing the representation.

In this case the functionality mirrors exactly that of the case discussed above where a representation of the record is supported by a separate 'package' of metadata designed to facilitate resource discovery and the implementation of the many other functions required to manage the record. Storing the representation within a database rather than simply within the hierarchical filestore of an operating system may mean that the user of that object benefits from the performance features, and the support for security, physical integrity and audit control which are typically an integral part of the database system (or of an application layer built on top of the database). This is the functionality typically provided by a document management system (where emphasis is frequently placed on the ability to control the integrity of the document within a multi-author environment).

However, it in no way changes the nature of the operations which can be performed on that object once it has been retrieved. Assuming that the object can be returned from the database in the same representational form in which it was stored, all the limitations described above in the context of working with presentation rather than structure continue to apply if that object is a representation which uses an encoding format predicated on the description of presentation rather than structure. (If the database system itself performs some sort of transformation of the document object into its own internal binary form for storage, then that does introduce a critical dependency on the capacity of that system to restore the object to its initial representational form.)

There are certainly benefits to be gained from the use of a database in this context, but they are in the domain of storing and querying the metadata properties associated with the representation. As noted in the previous section, that functionality could be implemented by storing the representation object itself outside the database.

6.3.2 Document stored as structured text in database

It is also possible to store documents in a database, not as a single object, but in a manner which attempts to describe or reflect some of their internal structure. It is sometimes argued that the table-based *Relational DataBase Management System* (RDBMS) model is poorly suited to the ordered hierarchical structural paradigm generally used to describe document structure (which is the model that underlies SGML and XML). It is almost certainly true that a complex hierarchical document structure, with deep nesting of elements, which are variable in terms of their size, and number of occurrences, is difficult to represent efficiently in this way.[23] However, if the structures can be simplified to those of tables, such an approach may be possible. Documents might be created using tables within a word processor tool or using a forms interface, and units of the document could be stored as fields in the database at the granularity of, perhaps, the paragraph.

The object database model dovetails better with the view of a document as an ordered hierarchy of component parts, and it is the claim of content management systems built on such tools that they can handle efficiently structured documents of arbitrary complexity.[24] However, such an approach presumes that such a complex structure has already been created by other means before the document reaches the database: this falls within the scope of the discussion of structured documents in section 6.4.

In this case, whether the database is object-oriented or relational, the metadata property set is available for structured querying, **and** the document content itself is held in a structured form and may be manipulated in more complex ways. This overcomes many of the limitations of the

[23] An approach to using an RDBMS to store XML documents of arbitrary complexity, originally suggested by Mark Birkbeck on XML-DEV, is outlined in Bourret http://www.informatik.tu-darmstadt.de/DVS1/staff/bourret/xml/XMLAndDatabases.htm

[24] See for example Poet Software, XML White Paper - XML Repository Requirements http://www.poet.com/products/cms/white_papers/xml/repository.html

'presentational' approach discussed in section 6.2 as the content is available for processing and reuse, though the flexibility available depends entirely on the granularity of structure which is described.

It would also be possible to associate metadata properties with component parts of the document rather than simply the document as a unit.

Another consideration in adopting this approach is that both metadata and document are held within the larger 'framework' of a proprietary database system and this entire system will need to be migrated at some point either to a new database system or to a different structured data storage environment. If a database system is to be discontinued (as a result of obsolete software or a change of policy) the information within the database needs to be exported and made available in some other way if the record is to survive, since the use of proprietary database software package is critical to accessing the information.

6.3.4 The database approach and the digital record

As discussed above, the performance of the database approach in terms of subsequent manipulation and re-use of digital records depends crucially on the granularity of information storage in the database system and the nature of the objects that it stores.

At one end of the spectrum, there is no real difference between the presentational approach of the previous section and at the other a very large amount of structural information may be stored in the system. In this latter case, the major worry with regard to longevity of the digital record is the continued ability to access the information as it is migrated form one version of the database system to another or from one database system to another as systems become obsolete.

A database type solution lends itself to metadata capture, to version and access control and to the facilitation of workflow. The potential for audit in record creation and use is also higher than that likely to be achieved in word processing based systems. The main problem that is likely to be encountered is that of finding a suitably flexible database model to cope with the complexity of the record. Forcing a record into an inappropriate model may restrict subsequent record manipulation and make re-use of the information within the record difficult.

6.4 The structured document approach

The approaches described in section 6.2 and section 6.3.1 place no requirement on the document creator to modify their authoring practices. They allow the creator to continue to create documents using whatever techniques of a word processing program they prefer. The additional data required to manage that document effectively is supplied separately, either by the creator or by other agents. The reusability of the document is conditional on the sophistication of subsequent processing applied to it, some of that may be highly specific to the creator's word processing environment.

The techniques discussed in this section adopt the premise that if the document creator can be persuaded to modify their practices and to adopt standardised approaches then they can contribute to making their documents into a more durable and flexible resource. The techniques discussed below represent a 'middle ground' between the use of the presentational approach and an approach based on databases. They attempt to make at least some aspects of the structure of the document explicit, and therefore more accessible to subsequent processes operating on the representation created.

6.4.1 Document structured using styles

The features such as the 'style' of *Microsoft Word* (discussed in section 6.2.1 above) and the related ability to store collections of styles outside documents in 'templates', were designed to facilitate consistency of formatting, within documents and across collections of document.

However, as noted in section 2.3.3.2, formatting is not applied to parts of a document arbitrarily, but rather in order to provide a means of identifying the logical components of the document. A document creator applies a particular *Word* style repeatedly to selected pieces of text because those pieces of text are structurally similar - they are of the same 'element type'. Through the

simple device of giving the style a name which corresponds to the element type (e.g. 'Section Heading') rather than just the presentational characteristics (e.g. 'Big and Bold'), then the document creator has taken a small step towards adding to the document useful data about logical structure.

A document which applies *Word* styles in order to identify logical structure rather than simply as a formatting shortcut will probably employ a larger number of named styles than a document which uses them as presentational shortcuts only. Section 6.2.3 mentioned three possible uses for italicised text, and the presentational effect could be achieved through the blanket application of a single 'Italic Text' style. However, to distinguish structurally the three different element types, three distinct styles would be needed, even though they would all be associated with exactly the same formatting characteristics.

The names of styles associated with sections of text are available not only to the author but also to processes operating on that text. If the styles are drawn from a predefined set - for example from a supplied set of templates designed for the creation of a number of document types - transformation processes operating on the representation have access to some description of the logical structure of the document, albeit at a fairly rudimentary level.

These styles may be recognised and processed in a variety of different ways:

- either within the document creation environment (*Word*) where some of them such as *Uniform Resource Locators* (URLs) may be hidden, but used at a later stage to form links from appropriate anchor text or where the document may be represented in different ways by applying different presentational forms of the *Word* stylesheet for different audiences or in different presentational environments.
- or outside the document creation environment when the styles used are recognised by software which can make use of this information in re-purposing the information as part of the construction of a finding aid (search index), transformation to different presentational representation (e.g. *Portable Document Format* (PDF), *HyperText Markup Language* (HTML), *PostScript* (PS), etc.) or the generation of different types of document from the information (e.g. action list, summary report, attendance list, etc. from a set of minutes).

There are many limitations of an approach, which relies on something like the 'styles' feature to communicate structural information.

- there are inherent limitations in the complexity of structure which can be conveyed by these means because styles can not be nested to reflect an implicit hierarchical structure. It may be possible to configure the processing application so that such nesting is implied by the names of styles; however, attempting to address more complex structures in this fashion leads rapidly to a proliferation of styles which are extremely difficult for the creator to manage.
- the imposition of restrictions on the use of styles - perhaps to 'validate' the document structure against a model for a document type - requires additional programming to perform those checks and may anyway be counter-productive.
- the nature of the word processing environment is such that the 'labelling' of the structural component remains 'tightly coupled' to the presentational properties associated with it in the word processor environment.
- the use of the familiar authoring environment may encourage a tendency for the author to perceive the presentational aspects as taking precedence over the structural ones, with the result that there is a temptation to adjust the presentation of the document directly rather than applying an appropriate style. This may produce results which are confusing to the author if, as is the case in the tools supplied by the ERM project, that author-supplied formatting is disregarded by downstream processing which applies presentation strictly on the basis of the logical structure.

This is the approach taken in the *Committee Document System* (CDocS) tools described in section 5.5 and appendices 1 and 2.

6.4.2 Document created using structured editor

Descriptive markup technologies are based on the idea that the description of logical structure should be quite separate from the description of presentation. To make use of such technologies, a structural model of a particular document type must be defined. This model describes the parts of the document and the relations of the parts one to another. *Standard Generalized Markup Language* (SGML) and *Extensible Markup Language* (XML) provide a syntax for describing the logical structure of types of document and for defining a descriptive markup language which can be applied to documents which conform to those structural models.

Formally, the document model is defined in a *Document Type Definition* (DTD). Working directly within a DTD is difficult unless the document creation/editing environment (Word Processing program, editor, etc.) helps the document creator to conform to the DTD and simplifies the choices to be made when applying structure to the document, by offering only those relevant in the current context. Structured editing tools allow an author to create documents in accordance with such a structural model which has been formalised as an SGML/XML DTD. As they create the document, the author identifies component parts of the text according to their logical function, and the editor software ensures that components are used in accordance with the restrictions expressed in the DTD. So, for example, the editor would not permit the creation of a chapter heading within a regular paragraph, as opposed to at the start of a chapter. It is this element of 'enforcement' which sets SGML/XML editors apart from word processing programs and which whilst allowing their users to create fairly rich structured documents, are happy if the user departs from the intended structure or simply ignores it altogether.

The SGML standard (ISO 8879) was defined well over a decade ago and provides a very rich syntax for the definition of document types. It is however this very richness which has been a stumbling block to its widespread adoption. The scope for markup minimisation and other short-cuts, has made it very difficult to write flexible editors with intuitive *Graphical User Interfaces* (GUI). The development of a more tightly defined, web-optimised version of SGML - XML has lead to an explosion of tools for the creation and manipulation of XML and of much more widespread adoption.

While an SGML or XML based approach to document creation satisfies some of the technical requirements for the creation of durable, reusable digital documents, it has the disadvantage that few clerical staff (or any other class of staff for that matter) are familiar with such tools, or with such an approach to creating documents. Such tools may be rather less well integrated with the standard work environment than, say, a desktop word processing package.

6.5 Conclusions

The different document creation approaches and techniques may be assessed according to several different 'axes'.

The first of these concerns the amount of change in working practice and thus the extent to which training and education is needed for document creators. The issue here is whether a digital record keeping system accepts whatever the document creators are used to doing and either uses back-end processes to try to fix any features of current working practice which create problems for records management or ignores the deficiencies, to the possible detriment of the record system. At the extremes are those environments which require no change of working practice and those which require document creators to use a completely different set of tools to those they are used to and adopt a radically different set of working practices. In the former case there is no training load but a considerable investment required in back-end processes and systems, whilst in the latter there is a major training investment required and an investment in new tools and systems.

The second concerns the means of access to the storage system within which representations are held as digital objects, and the extent to which that means of access introduces another layer of dependencies on the tools of individual software vendors. At one end of the spectrum there are systems based on organising these objects (which may of course themselves employ either proprietary or standard encoding formats) within the directory structures of the operating systems which are commonly in use. The only tools required to access the objects are those

normally used to process representations held in that encoding format (word processing, spreadsheet, browser, editing, etc. programs). At the other end of the spectrum are storage systems which 'ingest' information into an environment which requires the system itself to access that information. These latter systems may provide greater control over managing the representations and their associated contextual descriptive data but they introduce a another significant dependency on proprietary software, and so are more vulnerable to factors outside the organisation's control. If the storage system becomes unavailable in the future, there is a real risk to the accessibility of the objects stored within it.

The essential difference between the three 'approaches' to record creation described above is that they each privilege the capture and storage or preservation of certain facets or properties of the digital document over others. Some emphasise the capture of structural description over presentational data. The explicit capture of contextual description is an element that is significant whichever approach is adopted. It is, however, a consideration frequently overlooked in the preparation of documents intended only for rendition on paper, and the implementation of techniques to address this is critical in each of the cases discussed above.

As a result of these decisions to include or exclude properties, each approach may result in limitations on the operations that might be performed on the digital object at a later stage. The extent to which these limitations are significant can only be assessed in the context of the use requirements for the digital object - and in our case, **to the requirements of the digital object as a record**.

The ERM project wished to exploit the benefits of reusability which accrue from a structured approach to document creation. It was recognised, however, that it was not feasible to demand that a large number of record creators of varying skills and backgrounds should adopt the use of a new set of software tools specifically for this purpose.

The tools described in section 5.5 represent an application of the approach suggested in section 6.4.1, where the features of a familiar desktop word processor are exploited firstly to emphasise the description of the structure of the document as well as to control formatting, and secondly to capture the values of a small number of essential metadata properties associated with the document.

It might be seen as a 'middle path' between an approach which allows users to continue focusing exclusively on presentation for print and one which shifts them to a completely new tool set (whether that is a forms front-end to a database or a structured text editor). And although they continue to create in the first instance a representational form which employs proprietary encoding formats (with all the proprietary software dependencies which that implies), the characteristics of that form - the availability of some description of structure - are such that it can provide the basis for transformation to other more flexible and durable representational forms.

However, as has been discussed, there are limits to the structural complexity which a styles-based approach to structuring can address: a word processor can never be a substitute for a structured editor. Furthermore, in practice, the project found that some documents originated from external sources where the CDocS tools could not be deployed or had structures which the CDocS models could not easily accommodate. In such cases, a 'presentational approach' was adopted, with the emphasis on generating HTML-encoded representations which reproduce more or less the presentational characteristics of the word processed forms in order to meet the short-term requirements for record distribution.

7. The Purpose and Use of Digital Records

7.1 Introduction

A prospective user of a digital record must first be able to determine the existence of the record; secondly they must be able to establish a location from which to retrieve a representation of the record; and thirdly they must be able to make a request for the representation from that location. Any of these operations may be subject to authorisation restrictions.

At some stage in this discovery and retrieval process, users must also obtain sufficient information about the characteristics of the representation to allow them to make effective use of it. Such information might include a description of the encoding format and the program (or class of program) required to process the representation, but it might also extend to information required to make correct interpretation of the semantic content of the record. Both of these categories of descriptive data should form part of the metadata associated with the record.

Assuming that the user has successfully gained access to the information content of the record, the purposes to which they put that content are, of course, quite impossible to limit or predict. It may be possible to forecast with some confidence some of the operations which might be performed during the immediate period of active use of the record, but it is quite possible that at some point in its existence a record will be used for purposes quite different from those for which it was created. For example receipts for transfer of money by bankers in the eighteenth century have been used to establish the itineraries of grand tourists.

In order to enable this 'end user' activity, the management system itself must perform various operations on the representation objects and their associated descriptive data. These operations may - perhaps should - be more or less invisible to the end user of the record, but they are a critical component of the management system.

Given that the particular characteristic of the record is that it serves as evidence of an activity, and given the nature of the use contexts in which that aspect of the record is inspected most closely, a significant subset of the 'uses' of the record are concerned with testing claims and counter-claims surrounding its integrity and authenticity.

One final point to note is that the 'users' of the record will increasingly include not only human users but external software processes where rights of access must be strictly controlled.

7.2 Operations performed by 'end users'

The uses which might be made of a document-based record vary enormously depending on the nature of the record content.

7.2.1 Sharing the record

Whatever uses may be made of the record in the longer term, the primary purpose for its creation is to communicate information from its creator to a known user constituency. Depending on the nature of that information, there may be a corresponding requirement to limit access to members of that constituency.

7.2.1.1 Dissemination

There are two broad models of information distribution

- a 'push' model, in which the creator or distributor actively sends information to the user constituency
- a 'pull' model, in which the responsibility lies with the user to locate and retrieve the information

In this form, both models are problematic: over-reliance on 'push' can result in creators distributing information to users who in practice have little interest in the content, which itself

can result in information overload; while a dependence on 'pull' suffers from the difficulty that potential users need to be aware of the existence of information before they can begin to locate it and retrieve it and may result in disenfranchisement.

In practice, dissemination often employs a hybrid approach. For example, the creator sends out to users, not a representation of the object itself, but some notification of its existence, location and availability, and the users must act on receipt of that notification to obtain access to a representation of the information itself. Care is still required to ensure that those notifications are targeted appropriately, which is one of the motivations behind the idea of information subscription, where individuals register an interest in a type of information and (providing they have valid authorisation) are notified when items of information within that area are added or updated.

This is broadly the model for the distribution of the committee papers and their use by committee members, with the exception that the 'subscription list' is a controlled one: committee members are by default registered as 'subscribers' and subscription is not generally available to non-members. The clerk transfers representations to the Intranet server and notifies members by email of their location on that server (or more likely, of an index page which in turn points to the representations of the records.)

The selection of an appropriate dissemination strategy (i.e. establishing the right balance of 'push' and 'pull') depends on factors specific to the context: the nature of the information, the character and size of the user community, the means of distribution available, and the frequency of updates.

7.2.1.2 Access

Access to information is increasingly regulated, obviously by the data protection act and the convention on human rights, less obviously by other legislation and conventions. Institutions must be able to demonstrate that in both the printed and digital order access cannot be obtained to data without authority and that when such information is made available even to a closed community the owners of the information have given the necessary consent. For example it is illegal to divulge home addresses or telephone numbers unless individuals have given their express consent. Because of the openness of the digital order it is important that the institution's compliance officer is satisfied that adequate procedures and practices are in place.

7.2.2 Discovering the record

An authorised user must first be able to determine the existence of a record and establish the location of its digital representations. This information may be supplied directly to the user by a process such as that described above; in other circumstances, the user has to perform a process of 'discovery'. Such a process usually involves a user query requesting information about the availability of resources for which one or more metadata properties have certain values. e.g.

- which documents have **John Smith** as **author**?
- show documents of the **committee** called **Standards Working Group** for the **meeting date** of **1 April 2000**
- which items cover the **subject** of **digital signatures**?

The capacity of a management system to handle such queries hinges on the capture of metadata properties at a suitable level of detail, appropriate document structure and the construction of an indexed form of the metadata dataset against which queries can be run.

The property set which has been adopted by the ERM project to describe active records and their representations is drawn from:

- a subset of the Dublin Core Metadata Element Set[25]
- a subset of the Record-keeping Metadata Standard for Commonwealth Agencies[26]

[25] *Dublin Core Metadata Element Set, Version 1.1: Reference Description* (July 1999). Available at: http://purl.org/DC/documents/rec-dces-19990702.htm

(This approach is shown in detail in appendix 3).

Values are captured as part of the record creation and distribution process, and descriptions are stored as RDF-based XML-encoded documents. These descriptions provide a database describing the active records. That database supports both resource discovery and the management functions which must be performed on representations of the record to guarantee its preservation and continued access.

In response to their query, the user receives a (possibly empty) result set of those objects which meet their requirements, or, more likely, a result set made up of some descriptive/identifying properties which are sufficient for the user to make a selection.

In practice the development of a resource discovery tool which works on the metadata described above is a task which the ERM project has not addressed.

7.2.3 Viewing the record

Having performed a query, selected a (representation of a) record from the result set, and retrieved that representation from the specified location, a user generally wishes to view the content, and so requires a program which can process that representational form and render it in some output medium, usually on a screen display or in printed form, or both.

Many record users are accustomed to the fact that desktop operating systems and software packages typically conceal this chain of associations between some property of a digital object and a program to process the object. Indeed they become aware of its existence only when it fails - when, for example, they receive a file and the desktop operating system is not configured to associate it with a suitable program.

When we begin to consider the longer term accessibility of the record, however, we face the challenge of supporting a wider range of representational forms and software to process those forms on a range of platforms. At any one point in time, those associations should be sustained in a manner which is as transparent as possible to the user of the record - and this is the case for the committee papers distributed via the Intranet where the association is made by the browser program on the basis of *Multipurpose Internet Mail Extensions* (MIME) type. The longer term management of these associations between representational form and rendition software is a critical part of the preservation process.[27]

7.2.4 Referencing the record

The creators of other information resources need the ability to refer to existing records, and possibly to component parts of records and to aggregations of records.

This is frequently done through the use of informal and unstructured forms of reference convention ('the report of the standards committee', 'the minutes of the previous meeting'). However such forms function effectively only because the 'scope' of the reference - the range of documents from which the specific target is to be identified - has already been limited in some way and because the human reader of the reference adds additional information drawn from the context of its use in order to determine its intended target.

If a computer program, be it a mechanism performing some process on the documents or simply a document viewer such as a web browser, is to be able to resolve such a reference effectively, then a greater degree of precision is required.

In order for a collection of documents, on whatever medium it is stored, to be useful as an information base, units of information require clear, consistent and unambiguous identifiers which can be used subsequently to refer to those units of information.

A clear distinction needs to be maintained here between the identification of the information units and the names given to the physical objects which constitute digital representations of

---

[26] National Archives of Australia, *Recordkeeping Metadata Standard for Commonwealth Agencies* (National Archives of Australia, 1999). Available at http://www.naa.gov.au/recordkeeping/control/rkms/summary.htm

[27] See Chapter 8 for further discussion.

those units. The latter is dependent on characteristics of the representational form, particularly those characteristics related to physical structure, and on characteristics of the system within which that representation is stored. The same record may take the form of a single *Microsoft Word* document or a set of several hundred *HyperText Markup Language* (HTML) documents. That set of HTML documents might be stored as a single bounded unit in the form of a 'zipped' file but not necessarily if the metadata provides the mechanism for joining them together logically. And any of these representations might be stored as units within a database management system. The 'names' used to manipulate these physical objects are context-dependent, quite possibly temporary (since physical storage structures are liable to change) and irrelevant to the end user.

The management system must, of course, incorporate a mechanism by which the use of a 'logical identifier' can be resolved into a pointer to the current physical location of the representation(s) of the record.

The ERM project has established a convention for the construction of a record identifier, on the principles that it should be easy for both a human author to construct and use and it should carry an indication of the origination and purpose of the record. The record creation tools incorporate the facility to generate identifiers for records on this basis. The project has not, however, addressed the implementation of a resolution mechanism, with the consequence that this identifier has not been used in the construction of references to existing records.

7.2.5 Navigating the record

It was argued in section 2.3.3 that documents are composite objects, and that they have structural components which a user manipulates.

Various document viewers, most notably web browser programs make use of hypertext functionality to provide navigational features. This may be as simple as the generation of a table of contents or the resolution of cross-references, but it might extend to the traversal of a quite complex network of relationships between parts of the document and between other external entities.

The ERM project has done little to develop this aspect of document delivery for the committee papers. However the adoption of a structured approach to document creation provides a basis for doing so. The documents are not 'closed' objects, which can be manipulated only as units within themselves, and are intended primarily for display; they are structured dynamic information resources.

Furthermore, the association with contextual metadata, which establishes relationships with other records, means that each of these structured resources has the potential to be part of a larger digital information base.

7.2.6 Analysing the record

Conceived in this manner, the content of a set of digital documents becomes available for processing in any way an application program may choose.

The complexity of such processing is restricted by the level of structural description applied to the documents. In the case of the ERM document creation tools, this is quite limited. It is sufficient, however, to make possible processes such as

- analyses of meeting attendance
- identification of members to whom actions are assigned, resulting in the triggering of reminder processes

7.3 Operations to support end-user activity

Although the ERM project has done a limited amount to support end-user activity, this remains a high priority in adding value to the digital resource. Such features will be developed as necessary, though these tools take time and effort to build. In the paper order the navigational aids were both inconsistent and capricious. There was no consistent system of internal

referencing and such indexes as were provided were remarkable for their idiosyncrasy and incongruity. For example, in one year bequests might be referenced under bequests and in another under endowments or the professor of history under professor rather than history and so on. The problem is self-evidently an information issue, which must be resolved by rules irrespective of the media. Rules, which must embrace names, places, subjects and so on, have been debated by information professionals since the first index was compiled. In the paper order it is possible to get away without systematically codifying or applying rules as reading a document can often provide the clue to the method (if any) of referencing. This is much more difficult in the digital order particularly in instances where a document only exists virtually and the route into it may only be by exploring links which are themselves dependent on some external scheme of reference. In the digital world there are two approaches to the problem - markup and metadata applied consistently from sets of given rules. In the future these will need to be used in conjunction.

7.3.1 Indexing

All indexes require rules. At the most trivial level Sir Hannibal Hammer should be found under Hammer not Sir or Hannibal and the University of Glasgow under Glasgow not University. For retrieval to be automatic in the digital order the whole name needs to be marked up as a bounded personal name but the elements within it differentiated so it is possible to tell the difference between title, first or given name and surname. This is not such a difficult task. The more complex it is to ensure consistency, however and to differentiate between individuals of the same name the more likely this is to happen. In larger organisations it is especially important to be able to associate an individual with a position and an office. In a dynamic record keeping system this will change over time, someone starts their career as a research assistant and may become a professor. There may be several professors of chemistry at one time and many more over time. There may be only one called the Regius Professor but there will be a succession over time and minute takers will need to modify their practices so that there is no ambiguity over identity, otherwise retrieval will be problematic.

Ideally the names employed in the procedures developed by the ERM project should be drawn from a central database or thesaurus. The ERM team was involved in discussions about the need to create such a database of people associated with the university. This is not as straightforward as it may seem. Defining the group is fraught with difficulty as not all the staff are in fact paid by the university, some work for the health authorities, some for associated institutions, some are retired, some are volunteers who serve on committees, and some work for other universities. All have certain rights and privileges and can claim to be members of the university. The finance office is only interested in maintaining records of those whom they pay, while the library needs to know who can legitimately borrow books. This issue remains unresolved but it is one any institution seeking to emulate Glasgow's example will need to address.

Other naming conventions are equally problematic, particularly units within departments or inter-disciplinary groupings. There is little consistency and just as with personal names there is no central authority. There is hope that this will be resolved as an outcome of a project to put the university calendar on line, but the resulting thesaurus will have to be dynamic as departments merge and units dissolve and coalesce.

Much more problematic is subject indexing, essential for retrieval but the terms are difficult to define. The way forward is hard to determine but as in other domains it will require a collective approach, particularly as many terms are dictated by external bodies, the funding councils, Universities UK (previously CVCP). Unlike names, subject terms cannot be marked up as they are unlikely to appear in the text with sufficient frequency to be located by search devices. It is more likely that they will be embedded in the markup procedures.

7.3.2 Work Flow

In the physical order it is relatively easy, if the files have been properly constructed and preserved, to determine who was responsible for drafting various elements within a document and therefore in shaping policy. This issue has taxed computing scientists for some time and sophisticated software has been developed to allow workflow to be tracked. Such facilities are a

feature of *Electronic Document Management systems* (EDMs) which have been reviewed for JISC by the INTRA project team at the University of Manchester[28]. The report makes much of this workflow capability, which it is claimed 'offers a significant opportunity for managers to both monitor work in progress and to see where resources would be best allocated in order to ensure a smooth flow of operations'[29]. As the report correctly suggests, such capability is best adapted to standard transactional records, for example student matriculations, claims for expenses and so on. It is less well suited, as that technology now stands, to the type of records that the ERM project addressed, committee minutes and so on. There are, however, text processing packages which do include such facilities. On the whole they are expensive to install and require common platforms and working practices (unusual in *Higher Education Institutions* (HEIs)). Moreover for most university minutes such elaborate procedures are probably unnecessary, unless there is some strong compliance requirement where policy development is as of equivalent interest to process (transactions).

The capability exists within the ERM approach to develop protocols to remind those who have been tasked by a meeting to carry out some specific instruction the outcomes of which they are required to report on a certain date. There is demand for such a facility and it will be implemented in the future. This information will allow the mapping of work flows and completion rates, increasingly necessary for effective project management and in determining the lifecycle of the associated documentation.

7.3.3 Transformation and Production of Representations

A digital record is created in one representation and transformed into other representations for use, each medium of display requires a different representation. Changing from one representation to another is known as transformation (a document created in *Microsoft Word* may be transformed into HTML). Different representations have different attributes. A document is transformed into each representation as appropriate: for delivery, storage, longevity, etc. An example of this is in the production of the university Calendar. It needs to be made available not just as a printed, bound copy or *Portable Document Format* (PDF) but also in navigable form on the *World Wide Web* (WWW). The source document (say *Extensible Markup Language* (XML)) from which future representations will be generated is not necessarily the original representation (which is probably *Microsoft Word*).

A display of a representation is a rendition of it, for example, an HTML page is the representation and how it looks on a particular browser is a rendition (or view) of that representation, a different browser may produce a very different rendition of the same representation, depending on the type and version of the browser and the settings which the user has chosen. To manage committee papers in the digital world, it is necessary to have various views of the documents available, for example, there may be several views of the minute: decisions or actions, minutes that do not show reserved business and the full text for authorised committee members. An example of how one piece of software has migrated a manual process to the digital world is *Microsoft PowerPoint* which offers a variety of views of the presentation: edit view of each slide, outline view of the presentation, speaker's notes, handouts and thumbnails.

7.3.4 Migration of Media and Format

In the physical order objects migrate regularly from one institution to another. For example an individual or family may keep their books and papers in their attic or strong room for generations. They may decide to deposit them in a library or archive or sell them. The library or archive may subsequently merge with another. The objective of such migration is preservation even when a sales takes place as nothing guarantees long term custody more than the price mechanism. In the paper order much of the migration is by happenstance whereas the same will not be true in the digital order where migration will need to be planned to ensure that

[28] *A Meta-Evaluation of Electronic Document Management Systems* University of Manchester (August 2000). http://www.man.ac.uk/intra/subproj/proj17.htm

[29] *A Meta-Evaluation of Electronic Document Management Systems* University of Manchester (August 2000), section 2.1.9. http://www.man.ac.uk/intra/subproj/proj17.htm

content is preserved as platforms are upgraded and operating systems evolve. There are significant issues bound up in such migration, which are not present in the paper order. As has been discussed the image of the document does not reflect the underlying content or structure that is simply a bit stream, for which there is no analogy in the paper order. However it may be important to preserve the representational form. For example it may be useful to know in a form how large the instructions were or the precise positioning of the elements as these may have an effect on user choice and could be the subject of subsequent litigation. The simplest solution to representational forms may be to preserve a physical rendition, but in many cases there is no physical rendition such as form only available on the WWW. In a court of law counsel will wish to establish from a representational form if it was the one that a witness saw or filled in. They certainly will never have seen the underlying database, let alone the structure of its relationship. This is a notoriously difficult area and one where counsel can easily demolish a defence even in the physical order. It is well known this is one of the reasons that the *National Health Service* (NHS) and the armed services have great difficulty in defending cases. What is needed is guidance from the courts in the shape of precedent but many lawyers are at present unfamiliar with the implications of the digital order for evidence. The best institutions can do is to second guess areas of sensitivity which may in an increasingly litigious society give rise to complaint and where it is cost effective to invest or at least explore the necessary migration infrastructure. In the case of HEIs the student record is an obvious area where both for business reasons and risk avoidance such investment is necessary. In any case HEIs will need to develop migration procedures for the large quantities of enduring academic information they generate and wish to retain and where a guarantee of integrity is essential[30].

7.3.5 Management of Metadata

Applying metadata at the point of creation is not sufficient in itself. Metadata is dynamic in both the paper and digital order. In the paper order cataloguing conventions change and as a result books are often allocated new references. The same is true for records. Collections of documents in an archive are often recatalogued as time and money allow. With changes in use different aspects of the content may acquire different significance. Functions and therefore naming conventions change and control vocabularies need to be updated. For example in most old established HEIs faculties of arts spawned faculties of social science in the 1960s and 1970s and users of their records need to know when and how this came about. When recataloguing librarians and archivists record old reference systems as these are likely to have been cited elsewhere. Moreover scholars keen to establish the authenticity of a book or document want to be certain about provenance. How did it come to be in such and such an archive, who were the previous owners, was the book or document part of a larger whole that has now been lost, and so on? *Mutatis mutand*i the same will be true for the digital order but more so. With changes in technology the data will need to be migrated to new platforms and this process will need to be documents for precisely the same reason along with checks on integrity and so on. Access rights may change, reflecting societal and legal constraints or the removal of them, associated documentation may be destroyed and so on. All this information will need to be included within the metadata. This process will become much more elaborate than before is because paradoxically in the digital order far more information can be stored and supposedly preserved, and this when coupled with the compliance environment demands far greater transparency. Not all institutions will be able to afford such elaboration but there are social and political pressures for any institution in the public domain to adopt such policies and procedures.

7.4 Testing claims of integrity and authenticity

Records are evidence of activity within an organisation. Their evidential quality depends on their integrity and authenticity, which may be established by technical and/or procedural means.

It is instructive to examine how integrity and authenticity are established in the paper world, before considering how these means may be mapped onto their digital equivalents.

[30] For discussion about rules of evidence see *Information technology law* ; Ian J. Lloyd (London : Butterworths, 2000 3rd ed.)pp.237-

7.4.1 Authenticity and Integrity in the Paper World

There are five facets to authenticity and integrity in the paper world[31]

- The Circumstances of Creation
- Evidence of Changes
- Who Keeps it[32]
- Where it is Kept
- Signatures, Seals, etc.

The first concerns the know facts about who created the record, when and where they did it and the purpose for which they did it. This information is essentially metadata concerning the record creation and some of it may be included as part of the record itself, whilst other elements may be found in other 'supporting' records. Records which have this documented provenance are more 'believable' than those which do not.

The second concerns finding clues as to changes or inconsistencies in the paper record. These may include such things as: crossing out, rubbing out, different ink or handwriting. In a paper original these may be readily discernible, but in a world of photocopiers, much of this evidence can be disguised. For this reason, photocopies of documents which are not counter-signed with a 'top-copy' signature are often not acceptable (e.g. wills, power of attorney declarations, etc.).

Traditionally, the weight that might be attached to a record would depend on both where the document had been stored and who had been its keeper. Such issues as who deposited the record with the keeper and who might have had access to it during it storage were very important. If a will was deposited with a solicitor or in a bank safety deposit box, where only the owner (and trusted third parties such as the solicitor) had access to it, then there might be less risk of a challenge to its authenticity than if it had been kept in an unlocked desk draw, accessible to anyone who visited the house.

The final element concerns signatures and seals. Signatures and seals are special 'devices' which are intended to do three things:-

- Assert identity - authentication
- Attest to it being unaltered - integrity
- Stop denial - non-repudiation

A signature is an endorsement, which (in theory), may only be put on a document by one person and is a unique 'mark' made by that person. Signatures on very important documents are often witnessed, to attest to the fact that the signature really does belong to the person concerned, but also that they have signed of their own free will and not under duress. A seal on the other hand involves the application of a unique physical object to the document. In theory the seal could be applied by anyone, but its security comes for the fact that it is stored in such a way that only the true 'user' can get access to it.

Placing a signature or seal on a document makes it very difficult to forge and if (in the case of a seal) it is also used to 'close' the document, it makes it very difficult to alter after it has been sealed.

---

[31] Jeff Rothenberg argues for a universal concept of authentication based on 'suitability for purpose' in his paper "Preserving Authentic Digital Information", in *Authenticity and Integrity in the Digital Environment* (Council on Library and Information Resources, May 2000). Available at http://www.clir.org/pubs/abstract/pub92abst.html

[32] See for example Clifford Lynch's article "Authenticity and Integrity in the Digital Environment: an Exploratory Analysis of the Central Role of Trust", in *Authenticity and Integrity in the Digital Environment* (Council on Library and Information Resources, May 2000). Available at http://www.clir.org/pubs/abstract/pub92abst.html

7.4.1 Authenticity and Integrity in the Digital World

There are considerable problems with records in the current digital world, as many of the long-established facets of the paper world have yet to be incorporated in digital form in the way that we work with digital documents.[33] The main problems are:

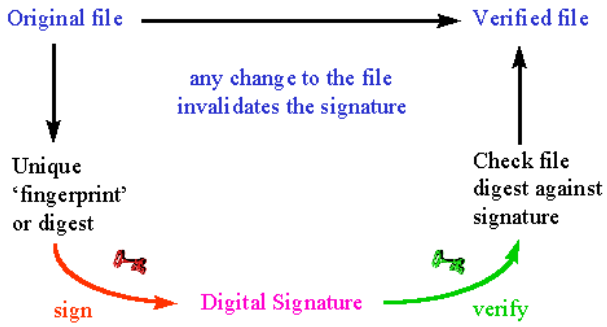
- The circumstances of creation rarely accompany the digital document (there is inadequate metadata about the creation)
- Changes in digital documents are very difficult, if not impossible, to detect
- Copies of digital documents are identical to the original[34]
- Digital signatures or seals are rarely used
- Date and time information attached to files in the form of modification date and time are easily falsified and therefore unreliable

7.4.3 Digital Signatures

Digital signatures are essentially an encryption of identity information, date and time information and a fingerprint (or digest) of the state of the document at that time. The last of these is very important because it allows one to check the integrity of the information contained in the record as the signature will be invalidated by any change to the file after it has been signed. It should be noted that a digital signature actually has more in common with a seal than a hand-written signature. Digital signatures employ encryption technologies with the following components:-

- a private key used to encrypt the signature information, this key belongs to the signatory and is kept securely by them
- a public key which is used to decrypt the signature information, this is made widely available so that anyone who wants to verify the signature and the integrity of the document may do so
- software to produce signatures from the document digest, identity information and date and time

A schematic diagram illustrating digital signatures



In a similar way to a seal, digital signatures employ a private 'device' called a key to which only the authorised user should have access and which is used to encrypt information concerning the identity of the signer and a digest of the state of the file being signed. The signature is verified

[33] These problems are well explored in Peter Hirtle's article "Authenticity in a Digital Environment", in *Authenticity and Integrity in the Digital Environment* (Council on Library and Information Resources, May 2000). Available at http://www.clir.org/pubs/abstract/pub92abst.html

[34] See David Levy's article "Where's Waldo? Reflections on Copies and Authenticity", in *Authenticity and Integrity in the Digital Environment* (Council on Library and Information Resources, May 2000). Available at http://www.clir.org/pubs/abstract/pub92abst.html

by decrypting with a public key which is made widely available. A good account of the technology involved is given in 'An Introduction top Cryptography' which is included as part of the free encryption and signature package 'Pretty Good Privacy' (PGP) [35].

For digital signatures to be widely adopted, the community using them needs:

- a proper environment for issuing the necessary keys, so that people accepting the signatures can have confidence that the 'keys' belong to who they say they do
- proper procedures as to who signs and under what circumstances
- verified date and time on the machines on which signing takes place
- a system for the storage of public 'keys' for verification of signatures
- software to verify digital signatures using the public 'keys'
- signatories must understand the need for safe storage of the private 'keys'

There are two significant problems with the use of digital signature:

- What exactly do we Sign?
- What do we do about Migration?

The first of these is relatively easy, it simply requires an informed decision. We can sign either a particular form (or all forms) of the document itself or a 'canonical' form of the document (that is a form that can be used to create all the other forms needed) it could be DOC, PDF, RTF, XML, HTML, etc. and the conversion/transformation process by which the other forms are produced from the canonical form (i.e. all the programs and control files).

The second is more problematic and will be discussed further in Chapter 8 under the general heading of migration in the context of long-term retention. In summary there is a simple choice:-

- Find the original signatory and get (s)he to sign the migrated version after carefully checking that they are the same
- Get a 'custodian' to verify the original signature prior to migration and then sign the migrated version to say that it is the same as the original (after careful checking)

Given that even the best migration strategies will result in a small amount of change in at least the layout and presentation of a digital record, the main question is how much difference can be tolerated after the migration?

7.4.4 The Way Ahead for Digital Integrity and Authenticity

Essentially the means that have been used for centuries to establish authenticity and integrity in the paper world can, for the most part be adapted to the digital world.

- The Circumstances of Creation - include adequate metadata with each digital document at the time of creation, which gives the necessary information concerning the circumstances of its creation (who, when, why, etc.).
- Evidence of Changes - the only solution to this problem is to employ digital signatures which attest to the state of a document at a particular time. If the document has been changed, it will not be possible to tell exactly how it has changed, simply that it has been changed and cannot be relied on.
- Who Keeps it - documents of record must be entrusted to a 'trusted custodian' or third party who can be relied on to keep the record where only those who have legitimate business accessing it can do so.
- Where it is Kept - the custodian needs to run a secure well managed repository where security, access and ensuring that technological progress and media failure does not render the record inaccessible are all taken care of.

[35] *Pretty Good Privacy* http://www.pgpi.org/

- Signatures, Seals, etc. should be applied at deposit, with validation and resigning at each migration to give added confidence in the authenticity and integrity of the record.

8. The Management and Preservation of Digital Documents

8.1 Introduction

The creation of a digital archive was not a fundamental part of the remit of the ERM project. However, one of the implications of the previous two chapters is that decisions about the creation of any digital object can only be made in the context of what uses will be made of that object. For the case of the digital record, at least some of these objects will be destined for permanent preservation, and that is another specific 'use' which conditions creation procedures.

This chapter argues that the preservation of a digital record is a complex process, some aspects of which are still not well understood, and that effective preservation strategies carry considerable resource implications. This is not to suggest that the preservation of paper records is cost free. However, the nature of the digital environment means that the locus of those costs may be different for the digital record and there may be a corresponding shift in some of the roles and responsibilities associated with the agents involved.

It has been argued that the circumstances of creation of a representation of a digital record are critical to its preservation. While this is true for the case of the paper record also, the 'preserver' of the digital record has a clear interest in shaping those circumstances to meet their requirements and gaining the co-operation of the 'creator' in that process. In some senses, the interests of the preserver might be regarded as different to those of the 'creator'/'user' - consider the case where the preserver wishes to ensure that complete contextual description/metadata is captured while the creator wishes to distribute the record to a user constituency as quickly as possible. One of the key challenges is to reconcile such conflicts.

The characteristics of the digital environment create a demand for a pro-active approach to the management of the record: doing nothing and hoping to recover later is no longer a viable strategy because of the high risk that access to the information content of an object will be lost within a short period of time. Furthermore, preservation in the digital environment is not a 'one-off' cost: it will require investments of resources for a potentially infinite period into the future.

8.2 How the digital record is different

Summarising briefly from chapter 2, the digital record differs from the paper record in the following fundamental ways:

- a user's 'experience' of any digital object is a complex product of the data itself and the operations performed on that data by software tools which enable it to be viewed
- the durability of digital objects is circumscribed by the limitations of storage media and the rapid change in hardware and software technologies
- a digital record may exist in many different representational forms each of which is appropriate for a particular use purpose
- almost any use of a digital record requires the making of a copy

As was noted above in Chapter 2, these factors have profound implications for how the digital record is managed (including how it is stored for long periods of time) and particularly how questions of the record's integrity and authenticity are addressed.

A digital object which is a representation of a record exists not as a single discrete unit but as an entity within a complex network of relationships:

- with the software agents which enable a human user to obtain access to the information content
- with other digital representations of the same record
- with other records (which may or may not have digital representations), or aggregates of records

- with the agents (individuals or organisations) who had a role in its creation and use, and with functions and activities performed by those agents

In order for the object to continue to have value as a record, the conservation and maintenance of these relationships is as important as that of the digital object itself.

8.3 Strategies for digital preservation

The strategies proposed to address the problem of digital preservation fall into three classes[36]:

- **technology preservation**: retaining the obsolete hardware and software technologies which provide access to the digital object
- **technology emulation:** developing new hardware and software to replicate the behaviour of obsolete technologies
- **migration:** the transfer of digital data to new technologies before the current ones become obsolete, in a manner which preserves the significant elements of content and functionality

The selection of an appropriate preservation strategy depends in part on the nature of the digital object.

The user's experience of a digital object is always 'mediated': it is the product of the data itself and the operations performed on that data by hardware and software tools. If it is thought possible to comprehend the 'original user experience' of the object, it must be decided how significant that experience is to the value of the object. In the physical world the 'look and feel' of a newspaper is valued by any user and cannot be replicated in other media. The experience of some classes of digital object is highly dependent on their processing. In the case of objects like games software, for example, the behaviour of the hardware platform is possibly as important as the content/structure of the data object. In such cases, it would seem as if only technology preservation or emulation can deliver this.

8.3.1 Selection of a suitable strategy

The strategy which is appropriate for one type of digital object or collection of digital objects may be totally inappropriate for another[37]. To arrive at a suitable strategy a careful evaluation of the experience of using the digital objects must be carried out.

One of the features of discussion preservation strategy in the literature is the rather absolutist stance taken by some authors. As in many academic debates this appears to stem from the fact that different authors come from different backgrounds/cultures and are addressing different issues. Appropriate technological solutions will differ depending on the nature of the digital object. For example, a multimedia representation will require an element of technology preservation coupled with emulation for it to work, whereas a database can be normally be migrated to a new platform.

At present there are no definitive solutions to which preservation strategies are appropriate. Even for the digital documents considered in this report.

8.3.2 A preservation strategy for digital documents

For the case of the record in the form of a digital document, it has already been argued that the experience of the record by two individuals with the same software tool is unlikely to be identical, as a result of different settings and different hardware components (screen resolutions, printers, etc.). As a result, the creator of a record cannot know the experience of the record

[36] It is not the primary purpose of this study to evaluate the relative merits of these strategies. The migration approach is generally considered to be the most promising, but it is acknowledged that the complexity and costs, and indeed the risks, of the process are variable and context-dependent - and, to a large extent, still unknown for the longer term.

[37] See Seamus Ross *Changing trains at Wigan…*for a synthesis of the current situation.

which others have even at the time of creation, let alone the experience which people might have in the future with different versions of software and on different hardware.

In addition, even in the short term, it is likely that an individual instance **must** exist in multiple representational forms as required for specific uses (editing, printing, browsing, etc.) and each of these representations may be regarded as an equally valid form of the record.

That is: a representation of the document created using one software tool can be transformed into another representation for use by another software tool, without **significantly** affecting the user's 'experience' of the record in the two cases any more than the differing experience of users using the same tool. This factor suggests that migration is at least a **feasible** strategy for the preservation of the document.

It is still the case that there has been less work conducted on emulation than on migration, and at least for the case of simple digital documents a migration-based approach to preservation may be appropriate.

Migration is not the only strategy but for our purposes migration seems to be suitable at present for preserving digital documents.

8.4 Migration as a preservation strategy

If it is accepted that migration is a **feasible** preservation strategy for digital records, this allows one of the major technological problems to be addressed - the fact that technology is constantly changing and thus representational formats have a limited lifespan. This however only moves us one step along the road to a preservation strategy.

As indicated in the previous chapter, maintenance of the authenticity of digital records has much in common with the maintenance of authenticity of paper records. It should therefore come as no surprise that the same principle applies to preservation. The main differences come from the fact that all experience of a digital record is mediated through computer hardware and software rather than being a direct experience of the record as it is in the paper case. This would not in itself be a problem if these technologies were relatively static, but as is well known, the pace of technological change is very rapid and is likely to remain so into the future.

8.4.1 Common features

In the paper world, the archivist/records manager needs to establish:

- appropriate environmental controls
- security
- periodic integrity checking

It is unsurprising that in the digital world these are equally important. Additionally, since even small amounts of loss of integrity can render the digital object inaccessible, there is additionally a need for:

- backup

These features are required irrespective of the preservation strategy selected, they are common features of all strategies.

8.4.2 Aims of migration

Migration is required to:

- combat media degradation
- combat technological obsolescence,
- improve cost-effectiveness
- meet new use/access requirements

The first two are much discussed in the literature[38] and although both costly and infrequently implemented in organisations, are reasonably well understood. The aim is not simply to ensure that access is to the information contained within our documents in the future. Since preservation is an inherently costly activity, it is essential to add value to the records by virtue of the way in which they are curated and by facilitating new uses or modes of access that are not practical in a paper-based archive.

The idea of adding value to documents by migration is not new. It has long been argued that an appropriate strategy for database migration is to preserve the data in an accessible form, together with details as to how it was accessed, rather than preserve the system (or an emulation of it) which allowed user access. In the future new means of access may be developed which will allow use of the data in ways that were not possible when it was created. There are of course methodological problems associated with such re-purposing.

8.4.3 Types of migration

The four migration aims listed in the previous section translate into three types of activity:

- refreshment: copy to another media instance of same type, without altering bits of representation or associated descriptive data
- replication: copy to a media instance of a different type, without altering bits of representation or associated descriptive data
- transformation: process which generates a new representational form while attempting to preserve information content

These types of activity represent, to an extent, a scale of increasing risk to the integrity of the information contained in the record, which must be of concern to curators. Assuming that there has not already been media degradation, the first represents little if any risk to the content. In the second case there is a slightly increased risk. This is less so if all the contextual information is contained in the 'associated descriptive data' and there is no reliance on information content in either the file names or the filestore structure which may be modified by replication.

The final case will almost inevitably result in some information loss. Most people will have experienced, for example, a change in some aspect of a word processed document when transferring it from one version of a word processing program to another or from one package to another. The extent of change may be minor, but there is a risk that it will not be and an important part of a migration strategy is careful checking of, at least, a sample of files migrated.

Many discussions of digital preservation recommend representations encoded in accordance with standards such as SGML and XML for long term storage, on the grounds that they are not only free of the limitations and dependencies inherent in the use of proprietary encoding formats, but they are by their very nature less volatile/more durable and the requirement for risky migration is minimised. While both of these claims are true, it must be recognised that

- **No** standard format is guaranteed to be permanent: at some time it will be necessary to migrate a representation of a record from one long-term storage format to another
- Transformations **will** be required to generate other representations from the long-term storage format and any event the range of 'delivery formats' required for user access will vary through time

It should be remembered that in this context archivists mean by 'time' long periods extending over at least a thousand years.

8.5 A functional model of an archival system

Irrespective of the preservation strategy, the objects to be preserved need to be maintained within some sort of framework which provides both the features needed for maintenance of

[38] See DLM-Forum *Guidleines on best practices for using electronic information* 1997, chapter 5 and T.Hendley, *Comparison of methods and costs of digital preservation*, (London, The British Library, 1998).

authenticity and integrity discussed in the previous chapter and ensures that necessary migration tasks to counter media degradation and technological obsolescence discussed in the previous section are carried out.

Faced with the problem of 'archiving' the very large amount of digital and physical objects collected and produced by the US space programme, the Consultative Committee for Space Data Systems put together a Model for an Open Archival Information System, which has been either adopted or adapted by many organisations and research activities around the world.

8.5.1 Reference Model for an Open Archival Information System

The Open Archival Information System (OAIS) reference model[39] has been developed by the Consultative Committee for Space Data Systems and is a set of recommendations concerning space data systems. That is not to say that it is applicable only to digital systems, data or space exploration. It has sufficient generality to cover the 'archiving' of digital, paper and other physical artifacts. It arose out of the need to ensure 'permanent or indefinite long-term' preservation of the vast quantities of digital data collected by the space programme. The people concerned in developing it were not, on the whole, professional archivists, but they clearly understood the problems of long-term digital preservation against a background of rapid technological change.

The OAIS work has been acknowledged as a useful model by people from a wide range of disciplines and is currently going through the ISO standards approval process.

The reference model defines the range of functions to be performed by any archive and aims to provide a common terminology and conceptual framework. It does not however specify any details of how such a framework might be implemented.

The reference model has been adopted/adapted by the:-

- Networked European Deposit Library (NEDLIB)[40]- whose core metadata is based on it.
- The National Library of Australia[41]
- Harvard University - who have set up an archive for thesis material based on it.

In addition it has shaped the CEDARS 'Metadata for Digital Preservation' specification which has produced a practical implementation of an archive for long-term digital preservation.

8.5.2 OAIS concepts

The draft OAIS reference model is a fairly complex and lengthy document. This section outlines the basic concepts and the next how they fit together. Inevitably such a simplification will be imperfect, but the aim is to give a clear view of the ideas behind the model and how they relate to the work of the ERM project.

8.5.2.1 Information Object

An Information Object is composed of two components:-

---

[39] *Reference Model for an Open Archival Information System (OAIS)* Draft Recommendation for Space Data System Standards. Consultative Committee for Space Data Systems - CCSDS 650.0-R-1 RED BOOK (May 1999) http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf

[40] A good outline of the NEDLIB project and its approach is given in the article by Titia van der Werf-Davelaar in *D-Lib Magazine* (Volume 5 Number 9) (September 1999) to be found at:- http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html
The NEDLIB homepage is at:- http://www.nla.gov.au/padi/metafiles/resources/130.html

[41] The National Library of Australia digital preservation site provides a wealth of international information on digital preservation and related issues. at:- http://www.nla.gov.au/padi/

Data Object

This is the actual information that was the focus of the ERM project. In the case of a digital *Data Object*, this is the digital file. An example might be a *Microsoft Word* file containing the minutes of a particular meeting.

Representation Information

This is the information which is required to be able to access the *Data Object*. In the example above, this includes the fact that it is a *Microsoft Word 97* format file. The information must be sufficient to enable the data object to be interpreted and its content rendered in an intelligible form. In our case there must be a reference to either a full description of the file format (at the bit level) or a means of rendering it, which is available to users of the archive (for as long as the data object must be accessible).

Representation Information may involve 'indirection' in that the *Representation Information* actually stored with our *Microsoft Word* file needs to identify it as that type of file but may point elsewhere in the archive to the detailed description of the format which needs to be recorded only once for each file type. Initially the detailed description may indicate where a copy of the software may be found, but this may be changed later if that software is no longer available and other approaches need to be taken.

There is of course a recursive element to *Representation Information*. The *Representation Information* itself will be stored (as a *Data Object*) in some file format (perhaps as *ASCII* text) and there must be *Representation Information* for that also.

8.5.2.2 Information Object Relationships

The relationship between the *Data Object* and *Representation Information* may be summarised as:-

Data Object -- *interpreted by* --> **Representation Information** -- *yields* --> **Information Object**

8.5.2.3 Types of Information Object

Two types of *Information Object are considered*.

Content Information

An *Information Object* containing content information is perhaps the primary information of interest. The *Content Information* is an *Information Object* which therefore contains a *Data Object* (in our example, the minutes themselves) and their *Representation Information*.

Preservation Description Information

In order for the *Content Information* to make sense, it is necessary to have additional information about the content (minutes), which will enable readers in the future to understand their context and the degree of confidence they may have in the content. These 'metadata' are termed *Preservation Description Information* in the OAIS model. They provide information in four areas:-

- **Reference** - identifies what the content is - basic description and metadata
- **Provenance** - describes the creation environment of the content (who, why, when, where), and the management history from creation to archiving, etc.
- **Context** - describes the relationships that the content has with other content and organisational structures etc., so that users of the content can gain an understanding of where it fits in

- **Fixity** - describes the ways in which content may be verified and its authenticity established - through for example checksums or digital signatures.

The ERM metadata is designed to serve as *Preservation Description Information* for *Content Information* created in the ERM testbed system.

8.5.2.4 Information Packages

An *Information Package* is a container for

- *Content Information*
- *Preservation Description Information*

Packaging Information relates the *Content Information* and *Preservation Description Information* and provides the information necessary to identify where the actual files concerned are. The *Information Package* is the 'unit' which archival finding aids identify and which are then of interest to users.

8.5.2.5 Types of Information Package

Submission Information Package

The *Information Package,* which is deposited with the archive, will be in the format in which the producer or creator of the information holds it. The archive is likely to make certain stipulations regarding minimum standards of *Representation Information* (what form is the Data Object in?) and *Preservation Description Information* (details of description, creation environment, context and fixity) that are required as a condition of deposit.

Archival Information Package

The information stored in the archive is likely to be stored in a different arrangement to that of the submissions. The submissions may be single items submitted serially over time, whereas the *Archival Information Packages* may be aggregations of submissions. A single submission may be added to a number of *Archival Information Packages*. These decisions will be made according to the policies of the particular archive.

Dissemination Information Package

When a request for information is made to an archive, the materials required to meet the request must be assembled and prepared for the consumer/user as a *Dissemination Information Package* or Packages. These will be assembled from *Archival Information Packages* and may be constructed to exclude information to which the consumer in question has no right of access.

8.5.3 Reference Model for Open Archival Information System

The Open Archival Information System (OAIS) presents views of the archive (and archival process) at different levels. At its highest level, it may be viewed as a black box receiving content from producers and sending content to consumers. At this level it is of little interest what happens inside the box, except that the producers can make deposits, safe in the knowledge that their content will be looked after and consumers are presented with a variety of archival finding aids, so that they are able to locate the content of interest and can then be supplied with it (including information about it or the means to render it).

Inside the black box there are a number of processes, which transform the material received into an archival form, manage the archive and transform material into suitable outputs for users. These processes are:-

Ingest

- accept submissions from Producers (as *SIP*s)
- prepares submission for archival storage (transformation to *AIP*s)

Archival storage

- stores, maintains, retrieves archived objects (*AIP*s)

Data management

- populates, maintains, accesses descriptive data and administrative data

Administration

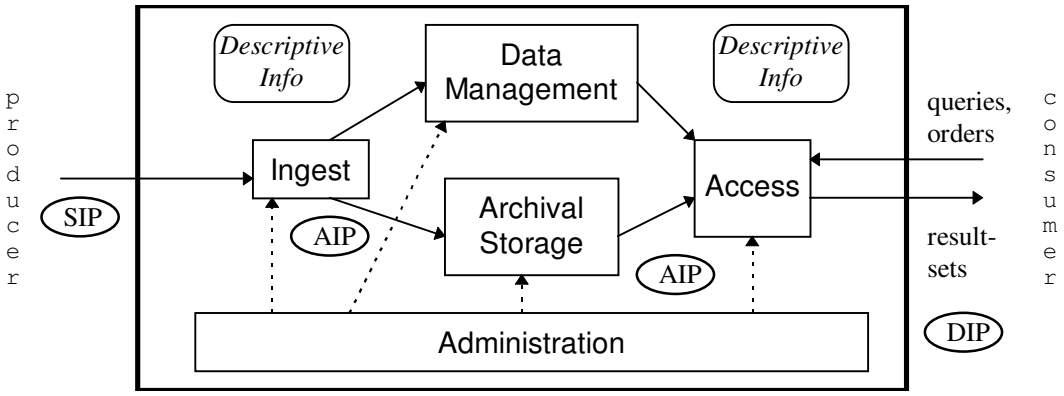
- general services to other functions

Access

- make information available to Consumers (via *DIP*s)

The relationship between these is best summarised in a diagram reproducing figure 4-1 in the OAIS draft..

figure 8.1 - OAIS Functional Entities



Note SIP (Submission Information Package), AIP (Archival Information Package), DIP (Dissemination Information Package)

Reviewers of OAIS model have requested that an explicit preservation function be added to the it, as this is an important function which needs to be a well planned process and is likely to require considerable resources.

8.5.4 OAIS, CEDARS, ERM

The ERM project has studied the OAIS model and the work of the CEDARS[42] project, particularly the metadata specification. What does this model mean for ERM in practice?

The OAIS model as described in the preceding sections provides a framework of what is required to construct a digital archive. Armed with the model one can test designs against it and explicitly address the components of the model or reject them as 'not required' in the application in question.

The CEDARS project is building on OAIS model to develop both examples of how it might be implemented and also detailed specifications of the metadata required for digital preservation (in the form of *Preservation Description Information* properties[43].

[42] The Cedars project can be found at:- http://www.leeds.ac.uk/cedars/
[43] The CEDARS paper "Metadata for Digital Preservation: the CEDARS project outline specification can be found at:- http://www.leeds.ac.uk/cedars/documents/Metadata/cedars.html

In the ERM project, considerable work has gone in to modelling the ingest process. There are two cases:

8.5.4.1 ERM Template Documents

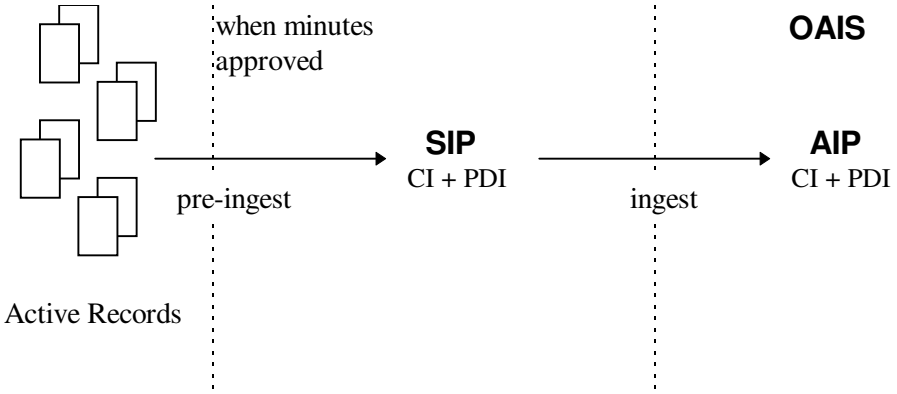
Documents produced in the *Microsoft Word* based creation environment for committee documents have a set of *Preservation Description Information* produced at creation time. The standard set of *Data Objects* produced by the upload transformations for these document types, described in chapter 5, can very simply have *Representation Information* sets added to form *Content Information Objects*. These taken together with the *Preservation Description Information* carried through the upload process and converted to an *XML RDF* object to form *Preservation Description Information Objects* form the basis for well described *Submission Information Packages*. The formation of the *Submission Information Packages* during document upload is termed 'pre-ingest' in figure 8.2 below.

There is little need for work on these *Submission Information Packages* before they can be accepted as *Archival Information Packages*, save the addition of fixity information.

Some work is needed to establish correspondences between ERM representations (outside archive) and OAIS objects (inside archive), but again this is of a relatively minor nature.

The project has demonstrated a mapping from ERM metadata packages (RDF) to the Cedars metadata schema using XSLT stylesheets to perform the transformation.

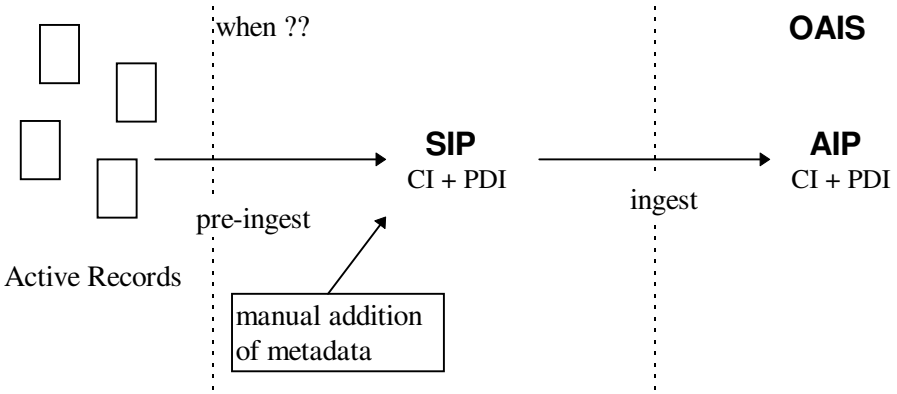
figure 8.2 - Ingest for documents created with ERM templates accompanied by DC/ERM metadata



8.5.4.2 Other Committee Documents

Some documents handled by the testbed committee system had to be accepted as they arrived from the body that submitted them to the committee, rather than being created or converted using the templates. With these documents the metadata required to satisfy the requirements of *Representation Information* is supplied by the upload process which uses standard transformations. The information requirements of the *Preservation Description Information Objects* need to be met by the person submitting the document supplying a basic set of metadata. It is then possible to combined this with the *Content Information* to form the *Submission Information Packages*. The formation of the *Submission Information Packages* during document upload is again termed 'pre-ingest' in figure 8.3 below.

figure 8.3 - Ingest for documents not created with ERM templates metadata added manually



8.6 Conclusions

Effective records management requires that attention is given to the organisation and preservation of the record. The processes involved are not well understood in the digital world, although the archive and records management professions have developed practice and procedure for managing these process in the paper world. It should be noted that 'archiving', in the sense in which computer people use it, is more concerned with placing digital objects in off-line storage than with the custodianship issues with which the archive profession concerns itself.

The digital record is different from paper in that whilst in the paper world, the record is experienced directly with the senses, in the digital the experience is always 'mediated' through the use of computer software. Since technology is constantly changing, there is a very real risk that the digital files will, without 'active management', become unuseable within a short space of time. A number of strategies are available for dealing with this 'creeping obsolescence', but none is without its difficulties.

Faced with these problems, the 'Space Industry' has developed the OAIS. The OAIS provides a functional model of what an archive should do. It does not provide any indication of how such a model might be implemented, nor the strategy to employ. This work does emphasise that much of the work associated with digital preservation is not about the technology, but about managing the processes involved. Well-managed processes don't come cheap and it has certainly been suggested that the archiving of digital material is rather more expensive than for paper. These ideas are neatly summarised by Clifford Lynch:-

> "The fundamentally hard things about managing bits into the future mostly aren't technical; they're economic and organisational. Bits need care and feeding. They don't do well with benign neglect. This means that we need to come up with financial models to keep these bits cared for and healthy as they are migrated into the future. We don't lose a lot of bits to technical failures in a well managed environment, but we lose a lot due to financial and organisational failures to maintain that well-managed, caring environment on a continual basis." Clifford Lynch (2000)

The CEDARS project has produced specifications, built on the OAIS model, of what is required in particular areas, particularly the area of metadata.

The ERM project has not implemented a digital archive, however the procedures it has implemented are shaped by the models and strategies discussed in this chapter, which are emerging to support that objective.

Glossary

archiving

An administrative procedural system established by an organisation to achieve economy, efficiency and effectiveness in the **selection, maintenance, preservation, access and use** of that portion of the records that has been selected or is designated for permanent preservation in other words 'forever'.

authentication

Is checking the assertion of identity, of persons or documents, to establish it is what it purports to be and has not been altered or corrupted at any time. This issue has both technological and procedural strands. 'An authentic record is one that can be proven to be what it purports to be, to have been created or sent by the person identified, and created or sent at the time purported' (ISO 15489-1:2001, 7.2.2).

authenticity

The condition of being authentic, trustworthy, or genuine. Attaining this in the paper order is dependent on procedures, involving where, how and by whom the paper record is stored, how it got there and who submitted it for storage.

compliance

Fulfilling official and legislative requirements, in this context the usual record keeping requirements relate to the Data Protection Act, Companies Act, Taxes Management Act, Freedom of Information.

curate (v)

Caring for an object, whether physical or digital, of historical significance in a managed environment forever. At present there is no agreement about preservation of certain objects for limited periods.

digital

When applied to information, documents, etc. - information stored in a form, based not on human readable symbols but on a binary encoding, which can be manipulated by computers (and thereby made readable by humans).

digital order

An environment that uses digital media as the output and the mechanism of record creation.

diplomatics

The study of the genesis, forms and transmission of archival documents; their relation to the facts represented in them and their relationship to their creator to identify, evaluate and communicate their true nature (Duranti 1998). It helped shape the legal theories of evidence, developed during the nineteenth century.

DSSSL

Document Style Semantics and Specification Language - is the international standard for the processing of SGML documents; mostly to format them for output to different media.

DTD

The part of the SGML document that defines its structure in terms of the elements and other structures that it contains.

evidence

In the archival sense can be defined as the passive insight into the processes, activities and events that led to their creation for legal, historical, archaeological and other purposes. The integrity of evidential value is demonstrated through the unbroken chain of custody (CLIR report - The Archival Paradigm).

finding aid

These are devices, such as subject guides and catalogues, which provide the key means of access to the collection. An inventory of record series within an archival collection is the principal finding aid as it combines arrangement information with that of content.

HTML

Hyper-Text Mark-up Language. A system for tagging various parts of a Web document that tells the Web client programs how to display the document's text, links, graphics and attached media.

integrity

Facilitates and ensures the ability to 'construct and maintain a history of intellectual dialog and to refer to that history over long periods of time' (Lynch 1994). In a digital environment this concern has two aspects - checking and certifying data integrity and identifying the intellectual qualities of information that make it authentic.

mark-up

A system (as HTML or SGML) for marking or tagging a document that indicates its logical structure (as paragraphs) and gives instructions for its layout on the page for electronic transmission and display.

medium

Material in or on which information can be represented, in either permanent or erasable form. For example, paper, microfilm, magnetic tape, magnetic disc, optical disc etc.

metadata

A term used for such a succinct description of the content of an information resource - a document, book, database, film and so on through the use of specific data elements.

migration

In this context it refers to the movement of data from one medium, or system, to another while maintaining the records' authenticity, integrity, reliability and usability.

paper order

An environment that uses paper as the output and possibly the medium of record creation.

provenance

An archival principle that aims to ensure that records remain, as much as possible, as they were originally created. In complex organisations it may involve viewing the business function through which a record came into being as the records provenance rather than the office or individual creating the record (CLIR report - The Archival Paradigm).

record

'information created, received and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business' (ISO 5489-1:2001)

rendition

It's the way in which information is presented to you, on screen or paper etc. It's conditioned by the material itself and the machine setting, software in use, output device, style sheets etc.

representation

The format used for the document / material e.g. RTF, XML, MS Word 97, LaTeX, TIF, JPG.

sealing register

A bound volume open to inspection in which entries of legally binding transactions (for example land sales and agreements) are recorded formally and verified by witnesses.

SGML

Standard Generalised Mark up Language. An ISO standard language for specifying the structure of documents and tagging various parts of a document.

standardisation

This has many meanings, some of which are mutually exclusive, in the migration of data from a paper based system to a digital system the main standardisation tends to be the normalisation of entities to permit retrieval, for example personal names and places. In a document management system, such as the ERM project was concerned with, standardisation usually means the creation of standardised formats and procedures. In the digital world it is essential that professionals from different backgrounds or perspectives have a clear understanding of standardisation in any given context.

transformation

Is conversion from one representation to another. It's important in migration and making information available on different media.

verification

In the physical world this means the substantiating the veracity of an element within an object which contributes to authentication. In the digital world it means ensuring that the object has a similar but more precise meaning. For example, it is possible to be reasonably certain that one written signature is the same as another whereas there is an imperative to ensure that one digital signature is identical to another.

version

In this context, it identifies a difference between the content of one instance of a document from another (usually bearing the same title). Usually it refers to the sequential production of a piece of work which is then numbered to indicate the progression through the creation and revision processes, e.g. version 1, version 1.1, version 2 etc.

version control

A process that allows for the precise placing of individual variants of documents within a continuum.

XML

Extensible Mark-up Language. A simplified form of SGML which allows the originators of documents (particularly on the WWW) to specify the structures in their documents and have these understood and displayed by browsers and other programs.