Appendix 4 The digital document deconstructed

1.1 The mediated experience of the digital record

Unlike a paper record where the symbols which encode the information are directly accessible to a reader, digital data is stored as a series of bits on a magnetic or optical storage medium - typically tape or disk - in a form which is **not visible or intelligible to the human viewer**. A machine is required to retrieve the binary patterns from the device, and then a program is required to interpret the encoding format of the bit stream and render it in a human-intelligible form. The information is generated as the result of the action of these hardware/software tools on the data.

Indeed the precise form in which the information is made available to a user depends as much on the action of these technological intermediaries as it does on the data on which they operate. Two different software tools may, for example, render the same data in ways that give the user different views (or 'experiences') of that data as information. This may be true even for the same tool running in different environments or with different parameters. Many word processing programmes incorporate multiple 'view' options which present quite different screen renditions of the same data. The user's experience of information becomes a complex product of the base data itself and the processes performed on that data.

This process of mediation also means that the user is largely unaware of **the physical structure** of a digital record - the way in which the bits of a representation are physically arranged on the storage medium. This will certainly change when a single representation is transferred from one medium to another, and perhaps on transfer from one instance to another of the same medium - it may even change as the operating system performs physical housekeeping operations on a storage medium.

1.2 The durability of the digital record

The storage media on which digital information is recorded have inherent characteristics that mean that **they degrade relatively quickly**, at least when compared with the known durability of paper.[1] To ensure continued access, the digital objects must be copied to a new medium (or to a new instance of the same medium) before the current one degenerates.

Furthermore, the technologies of storage media and machine tools which read those media and the encoding formats and the programs which interpret those formats, are **in a state of constant change**, driven by scientific advances, user requirements for improved cost effectiveness, and the commercial imperative of the suppliers.

The use of proprietary storage formats, the specifications of which are not in the public domain, creates dependencies on the software tools of individual software vendors. A file created using, say, *Microsoft Word*, and stored in *Word* format can only be read by another version of that same program, or by a program which is able to decipher the *Word* encoding format. If the file is transferred to an environment where such a tool is not available, or if it is not accessed over a period of time and during that time the program becomes obsolete, then the information content of that file becomes inaccessible.

This constant change is certain to continue: there is no evidence to suggest that a plateau of technological stability will ever be attained.

The strategies proposed to address this problem - all of which carry with them costs - fall into three classes[2]:

the preservation of the obsolete technologies themselves,

[1] Some written records have survived for over 3000 years.

[2] 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.

the emulation of obsolete technologies,

the migration of digital data to new technologies before the current ones become obsolete.

In short, these twin risks of physical volatility and of technological obsolescence mean that - whatever preservation strategy is adopted - the need for active management to ensure preservation is greater than for non-digital records. The lifespan of a (representation of a) record stored in a digital medium can be relatively short, and within that period records must be appraised and, for those identified as worthy of long-term preservation, action must be taken to ensure continued access to the information content of the record.

Indeed, it may become necessary to reassess notions of what is meant by 'the long-term'. If a record is to remain accessible, either as an 'active' record or as an archival one, for a period of time during which technological change has an impact then these issues become significant.

For paper, the lifespan of the record is determined principally by that of its physical medium, and that can be estimated, assuming that some basic environmental safeguards are in place. However, it is extremely difficult to predict with certainty a maximum period for which a digital record might be considered to be 'safe' from the threat of technological obsolescence. In contrast to the paper domain, the risk is not a simple function of the time elapsed since the record was created, but a product of many independent variables. In the worst case, access to the information content of a digital record may be at immediate risk if it happens to be created shortly before its encoding format becomes obsolete.

Faced with such a threat, it is tempting to suggest that one solution is to 'de-digitise', to create printed renditions of the records for the purposes of preservation. Leaving aside the costs of physical storage space which such a strategy implies, it is increasingly the case that there are elements - possibly significant elements - of the 'user experience' of the digital record which are lost when the record is rendered in printed form. For example, the printed rendition of a database with a Web interface will not reflect the users' experience of entering data and therefore it will be impossible for them to be certain from the printed rendition that this was 'the form' they filled in. This raises the important issue of surrogacy and the definition of what is meant by 'original' in the digital order.

1.3 The representation and the record

Quite apart from the prospect of migrating a record from one format to another to ensure continued accessibility through time, it is quite probable that, at any one point in time, the requirements for the use of a digital record may dictate that representations of that record exist in **multiple encoding formats**. This may be true even within the active life of a record. For example, a format which meets the requirements for the generation of high quality printed output might not be suitable as an object for a sophisticated automated analysis of its content and different access devices and output media may require different output formats. In the case of the preparation of text for publication the digital version will include complex markup to drive the typesetting process which is not only useless but causes confusion if the text is analysed digitally.

These requirements - for multiple representations at any one point in time, and for the migration between media and formats through time - have a more profound significance for the management of the digital record. The content of a paper document is recorded on a specific storage medium and is never separated from it. The management of the content has been bound to the management of that (initial, 'original') medium: that single 'representation' **is** the record (though there may be circumstances in which a surrogate is used in preference to that original).

In contrast, the digital record may be created on one device and medium, but it is almost certainly **transferred to many different physical devices and media** during its life. It may be **transformed into different encoding formats** for storage, and rendered to a user in still further media and via other intermediate formats.

Consider the example of a matriculation form. In the paper order, the completed form itself constitutes the permanent record. In the digital order, however, the digital object which is created through the immediate action of completing, say, a *HyperText Markup Language*

(HTML) form has only a transient existence: the HTML form itself serves only as an interface, a means of creating a digital representation - perhaps a record in a relational database - which will itself serve as the basis for other representations later.

None of these individual representations of a digital record is inherently 'better' than any of the others: each meets the requirements of a particular function or use at some time. There is no absolute requirement to make the record available for all time using the same encoding format in which it was initially created, or to continue to reproduce exactly the 'user experience' of all the manifestations in which it was distributed and used. The record is the information (including its context), not one specific representation or rendition of that information.

However, the representations are the real objects that must be managed. In the most general sense, the transformation of one representational form into another is carried out in order to make the information content more 'useful' - which may encompass a whole range of purposes, from increasing its durability, to improving the cost-effectiveness of its management, to enhancing access to the record's information content. All such transformation processes, however, carry with them the risk of loss of some part of the information content of the input object, or of the 'functionality' which is associated with it. In some cases, such loss may be considered acceptable for the context of use - for example, the transformation of a document from a word processor representational form which can be edited to a read-only delivery format may be considered perfectly acceptable - even desirable - if the use context is that of read-only exchange. It has also been argued that transformations can increase the information content of a record.[3]

This separation of the medium and the message, of representation and record, for the digital document means that the challenges involved in persuading others of the trustworthiness of a representation of a record are different from those encountered when dealing with a paper record. Because the paper record is permanently bound to its medium, such concerns tend to focus on the integrity and authenticity of that medium; in the digital domain, where information can be copied, transferred, transformed and manipulated so easily (and **indeed must be copied and transformed if it is to survive**), they must extend to the **processes** which generate the representations of the record. Although this is sometimes the case on paper (for example the notorious case of the bogus Hitler diaries), it is a last resort rather than the first.

1.4 The ubiquity of the copy

Almost any 'use' of a selected representation of a digital record involves the making of a copy in some way. If a user accesses a representation which is held on a networked server, a copy of that representation is transmitted to their client computer, and indeed multiple users could perform the same action simultaneously on that same representational form. 'In the digital world, I share with you a file that has the same properties as the file I have - the original, as it were. Now I have it, and you have it, too.'[4] Like the requirement for multiple representational forms, this is a significant and fundamental challenge to assumptions and expectations shaped by practice in the paper world. Indeed, it can be argued that there are no 'originals' - certainly no unique items - in the digital domain whereas in the paper order 'my' copy is 'my' copy and 'your' copy is 'yours'. Providing both versions are preserved, a third party can identify the contributions to the construction of the document. This process can clearly be seen at work in paper files which bring together various versions of documents with contributions that can be attributed to different hands. This is self-evidently not so in the digital order were documents are regularly passed digitally between individuals who make changes and alterations but often

[3] Lynch, Clifford, 'The Integrity of Digital Information: Mechanics and Definitional Issues', *Journal of the American Society for Information Science* 45(10): 737-744 (1994).
[4] Clifford Lynch, cited in Smith, Abby, , 'Authenticity in Perspective', 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

without any means of identification unless the necessary processes are in place. This has important ramifications for management of information in the digital order.[5]

1.5 Integrity and authenticity

A claim that an object has **integrity** is an assertion that it has not been corrupted over a period of time. Assertions of **authenticity** encompass a broader set of claims which support the proposition that the object is what it purports to be: they might include claims that it was created at a specified time by a named agent. A digital representation of a record will only be accepted as legitimate evidence of an activity if it can be shown with a high level of confidence that it is what it claims to be - it is authentic - and it has not been altered - it has integrity.

One proposed technological answer to the problem of checking the integrity of a representation of a record has been the application of 'digital signatures'[6]. Digital signatures employ encryption and public key infrastructure technologies to bring together three pieces of information:

some information derived algorithmically from a digital object (which for the case of a digital record would be a specific representation of that record), sometimes described as a hash or digest,

some information unique to an individual agent,

a time stamp.

This resulting composite unit - a digital signature - establishes an assertion of the state of the object by the agent applying the signature at a specific time. A user of the object, at some later point in time, can create their own digest from the object. Assuming they have access to the public key of the initial signatory they can compare this new digest with the previous digest that formed part of the digital signature: **any mismatch highlights the fact that the object has been altered since the initial digest was created**.[7]

This form of integrity check is based on the concept of testing copies of digital objects for 'bit-for-bit' equivalence. However, this becomes quite insufficient in the context of an approach to the management of the digital record that is based on the premise that different representations of that record will be created at different points in time. The digest, which is algorithmically derived from one representational form, can not be compared with that derived from another representation of the record because there is no bit-level equivalence between the two digital objects.

As suggested above in the discussion of transformation, the question that must be addressed is that of the 'intellectual integrity' of the record. In particular of how to verify the extent to which the multiple representational forms accurately convey to the user the **significant** information of the record as it was initially created. Lynch expresses this in terms of checking that different representations all capture the same 'abstract essence' of the record, and proposes an approach based on 'canonicalisation', by which a digital representation of this 'essence' is created and all other representations can be algorithmically validated against that.[8] He acknowledges, however,

[5] See 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

[6] Good general introductions to Pulbic Key Technologies are to be found at:www.iplanet.com/developer/docs/articles/security/pki.html and www.pgpi.org/doc/pgpintro/

[7] This brief discussion will leave aside the problem that the user's confidence in the outcome of this test can only be as high as their confidence in the initial signature itself: that they trust that this is the signature associated with the initial object by the creator or custodian (rather than a substitute), that the public key used to decrypt the signature does indeed belong to the individual who purports to own it (rather than the supplier of the substitute), that they trust the apparent value of the time stamp etc.

[8] Lynch, Clifford, '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

that not only are there practical difficulties in implementing such a verification procedure but also that there is an absence of clear agreement on what characteristics constitute the 'essence', the 'canonical form', since that set of characteristics varies depending on the class of object. This is well understood in the physical order in that there is wide agreement among scholars of the canonical forms of different objects, whether they be archives, printed books or images.

The capacity to test other claims of authenticity must also withstand the complexities arising from the multiple representation case. Rather than the simple assertion that 'the digital object A was created by agent X', the claim becomes 'this digital object B was created from digital object A by a process P and digital object A was created by agent X'. In this scenario, it becomes critical that the stored representations of a record are supported by descriptive data ('metadata') which details explicitly - amongst many other things - the relationships between the multiple representational forms.[9]

These issues are critical as much litigation (for example, that concerning negligence) centres on the integrity, authenticity and completeness of the record. If a claimant can challenge the 'evidence', then the defender's case will fail irrespective of whether 'negligence' has actually occurred. Increasingly, external bodies (for example the Financial Services Authority – FSA) have powers to audit the processes surrounding the creation and preservation of records as a guarantee of integrity.

1.5.1 Record content

Since the record is to serve as evidence of an activity, the information content must be a reliable reflection of that activity. The reader of a record must be able to have some confidence that this was the case at the time of record creation and continues to be the case. The degree to which a reader considers a record to be reliable depends largely on the trust they place in the procedural (not only technical) controls which conditioned the creation of the record and its subsequent use, storage and management.

In the paper order conventions have evolved over centuries to address this problem. A royal seal carries power but only gives greater authenticity than that of a bishop if the controls that surround its application and use provide greater authentication. Monarchs relied on chancellories out of which most archive services were to grow. Letters have a clearly recognisable form which reassures the reader that it is what it purports to be and a style and vocabulary which defines its purpose. In the committee papers (encompassed by the project) the same is largely true and the final agreed minute is customarily signed off by the convenor as a 'true' record. The minutes are then entrusted to the secretary whose duty it is to preserve them and prevent any subsequent alteration. Such documents are then usually locked up so only a select few have access to them and so, in theory, the opportunity for tampering is limited. The degree of this fiduciary protection will depend on the significance of the minute. For very important minutes security in the physical order can be elaborate with locked volumes stored in safes. The important element in this procedure is that authentication is provided by someone other than the creator, who should have been party to the event and acts as their recorder. The custodian has a fiduciary responsibility to protect the physical representation but not necessarily to verify the content. Replicating such procedures in the much less secure digital order is fraught with practical, technical and legal difficulties.

1.5.2 Record context

The value of a record, regardless of storage medium, derives not only from its information content but also from its context, from the circumstances of its creation and use. Archivists refer to this as the 'diplomatic' of the record and some information experts have adopted this

---

and Information Resources, May 2000), especially pp 37-38.
http://www.clir.org/pubs/abstract/pub92abst.html
Lynch, Clifford, 'Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information', *D-Lib Magazine* 5(9) (September 1999). Available at
http://www.dlib.org/dlib/september99/09lynch.html

[9] See section 2.3.2 below for further discussion.

term. The record serves as evidence only if (amongst other requirements) its context can be described. Much of this information deals with relationships, between the record (or the activity it describes) and individuals or organisational bodies, or between the record and organisational functions, or between records.

Much information regarding the creation and use context of a paper document is recorded explicitly, either as part of the document content itself or as additions to that content. It may be in the form of annotation or labelling of the record itself or as supplementary description held separately and linked to the record as part of the record-keeping system.

Some contextual information, however, is conveyed implicitly through aspects of physical structure (for example colour of paper may carry a significance which is not explicitly recorded in another way) and particularly through physical context (for example physical juxtaposition with other records). For a digital record, physical structure and physical context will no longer be a constant **and all contextual description must be recorded explicitly**.

For example, relationships between paper records may be created by bringing them into juxtaposition in a physical file. In the digital domain, such strategies might be mirrored in the location of individual records within a hierarchical file store directory, the structure of which is designed to correspond to the structure of the paper file system.

However, in the digital domain, the physical structure and physical context of a record is liable to change through time. It may be possible to formulate the requirement that the physical file store arrangement should be preserved. However, such a strategy presumes that a user viewing that physical arrangement at some point in the future (if indeed they view the arrangement as a whole at all) will interpret it in the same way as its creator intended. What is significant in this example is not really the directory structure itself, but the logical relationships which it is being used to convey. To preserve the full contextual description in the digital domain, such relationships must be recorded explicitly. In the paper world surrogates are often made of documents so that copies can be placed in the files to which they relate. This may be considered unnecessary in the digital world but in that case the document which will undoubtedly relate to different subjects and areas of activity must be extensively cross referenced if the logical relationship is to be preserved. For example a letter commissioning a new building will have implications for finance, planning, the customer, contractor and so on.

A cautionary note is required with regard to this example. It must be emphasised that it is only a **possibility** that physical arrangement within a filestore conveys information about context. It is not **necessarily** the case, and it would be quite incorrect to make the assumption that because two objects shared a physical location, they had some logical relationship.

A corollary of this is that logical relationships can be created between entities which are physically quite separate, and a record may be part of multiple logical relationships without the requirement for multiple copies placed in separate files or folders. Many commentators view this possibility as an advantage of the digital order. The existence of relationships in the digital domain may be conveyed by techniques quite other than physical arrangement. For example, an HTML document may present a list of hypertext links to a set of documents, which are scattered across many physically separate servers, in a manner designed to communicate to the reader the existence of relationships between those resources.

Of course, such contextual information may be - and often is - lost in the course of the life of paper records unless it is explicitly captured, but in the digital domain, the risks are increased by the fact that physical structure and context are so much more volatile.

1.5.3 Contextual description and metadata

This contextual description is a small subset of what is broadly described as **metadata**, information about the data. Many discussions of metadata, particularly in relation to the World Wide Web, concentrate exclusively on its role in 'resource discovery', that is, being able to locate information of interest. However, the class of contextual information described above, while potentially useful to aid discovery of a record, serves also to provide a more complete understanding of that record and its use once it has been located. Indeed, several different

classes of metadata are required to perform the many functions associated with the management and preservation of the digital record, from the physical management of its representations (to answer questions such as, 'what program is required to view this object?'), to the documentation of the history of the record and its representations (when it was created, stored, transformed etc.) through to indications of access and use rights and restrictions.

The set of metadata properties required depends on the functions to be performed. The values of metadata are assigned at different points in time: some will be assigned at record creation and will remain associated with the record from that time on; additional properties will be assigned, or the values of existing properties amended or extended, to record significant actions performed on the record.[10]

In the paper order, failure to capture metadata explicitly at record creation and to retain it can perhaps be compensated for by the work of the records manager/archivist to uncover and record that context retrospectively. The fact that in the digital order so much of that contextual information is 'implicit' makes such a task potentially more difficult in the case of the digital record. For example when an email attachment is read the software package usually arbitrarily ascribes a filename which has nothing to do with either the original filename or the content. If however the document is received with embedded metadata then this problem is obviated.

Metadata may be embedded in a representation of a record or recorded separately and associated with it. Context of use may dictate that some metadata properties are both embedded within a representation and stored separately. For example, if an institution's Intranet server is indexed by a program which can exploit the presence of structured metadata property values within the 'meta' elements of HTML documents, it is eminently sensible to ensure that those properties (which would be useful for resource discovery purposes) are embedded in the HTML representations of the record. However, the fact that metadata must be preserved throughout the transformations to which the record may be subjected dictates that such metadata should be maintained as a separate but related digital object. The long-term management of the metadata object itself may require that it undergoes transformations from one representational form to another, independently and asynchronously of any transformations of the record and representations which the metadata describes. Understandably these processes are both complex and consequently expensive and the Records Manager needs to be aware of the cost / benefits.

1.5.4 Record structure

The structure of a record is the way in which the component parts are brought together to form a whole. A description of structure is concerned with distinguishing and with establishing relationships between those parts.

1.5.4.1 Physical and logical structure

Documents are composite artefacts, parts of which both the author and the reader identify and manipulate. In the paper domain, perhaps the most easily identifiable structure is that a document consists of a number of pages, or a series of volumes each containing a number of pages, but this 'physical' view of structure makes sense only in terms of a rendition of the document in one particular medium. By contrast the 'real' physical structure of a digital record - the way in which the bits of a representation are physically arranged on the storage medium - is irrelevant to the reader.

Another 'level' of physical structure is that physical structure which is manifested to the reader when the document is rendered. This physical structure is dependent on the format or medium in which the document is rendered to the reader. For example, the notion of the 'page' ceases to

[10] Establishing precisely what set of metadata properties is necessary for the effective management and long-term preservation of, and continued access to, a digital record is a considerable challenge. The ERM project has made use of two metadata schemas: the *Recordkeeping Metadata Standard for Commonwealth Agencies* (National Archives of Australia, 1999), and the Cedars project's recently published proposal for a generic metadata framework for preservation *Metadata for Digital Preservation : The Cedars Project Outline Specification* (Cedars, 2000), intended to support the functional requirements described by the *Reference Model for an Open Archival Information System*

have meaning when a document is delivered on the screen as a single continuous scrollable window, as is the case with the Web browser display of an HTML representation, and screen display features such as HTML 'frames' may permit the concurrent display of multiple component parts of a document in a manner which does not have a direct analogue in the paper rendition.

Although the physical structure of the record may be contingent on the context of use, the record has a 'logical' structure that remains constant across these different renditions.

Logical structure - often of considerable complexity - clearly exists in documents: a book may be divided into chapters and sections, an article into sections, which in turn contain sub-sections, paragraphs, lists, captions, footnotes, and so on. But most documents created using tools such as word processors do not contain explicit descriptions of that structure. There is usually no label attached to a piece of text stating that it is a title. In spite of this, authors are able to communicate complex information about structure to human readers of a document.

1.5.4.2 Structure and presentation

The key to this process is the use of a commonly understood set of **presentational** conventions to convey cues about structure. The authors apply formatting to the text not at arbitrary points, but in order to differentiate structurally distinct component parts of the document (and to establish correspondences between structurally similar parts). They make two decisions: first they classify the piece of text as an element type (section heading, author name etc.), and then they select and apply formatting appropriate to that element type in accordance with an implicit or explicit set of presentational conventions.

These conventions are dictated by context - a range of factors that might include the constraints of an authoring tool, the requirements of a delivery medium, the guidelines of a publisher, 'tradition', even personal preference. For example, a poorly sighted reader might prefer to represent all documents in a large font size and a traditional typesetter will never use half points. Some authoring packages automatically capitalise after a full stop even if it makes no sense to do so for example in abbreviations (e.g.). In such cases even if the authoring package has been set to ignore such conventions subsequent transformations may reinstate them. The reader actually performs a complex interpretation to establish information about structure that the author wishes to communicate from the presentational cues that they employ. If the cues are misused (or the reader does not share a common understanding of the author's conventions), or they are corrupted at some point in the course of the transmission of the document between author and reader, then there is the possibility that the structure is not communicated as intended.

The use of presentational conventions is related to the medium in which the document is rendered, and perhaps even to the functionality of a class of software tools used to perform rendition to that medium. Some presentational conventions make explicit use of physical characteristics of a medium, such as the insertion of 'hard' page breaks to separate logical divisions of a document. As noted above, a page makes little sense in terms of a continuous scrollable display; while a columnar text layout may be appropriate in a printed 'newspaper'-style rendition, it would seem unusual - inconvenient even - on a small screen display; and the difference in font sizes of headings and text may be rather more pronounced for screen display than for print. It could certainly be argued that the rendition of cross-references in the form of hypertext links which can be traversed by user action creates a significantly different experience from a printed footnote containing a 'see also' note.

The reader's 'experience' of a digital record is a product not only of the content of a representation of that record but also of the action of a software tool on that representation, and the result of that action is conditioned by parameters whose values may be beyond the control of the author or distributor of the record. In the case of a typical networked client-server environment, the author can make available a representation of a record on a server. If the server is available only to a defined user constituency (as may be the case for an organisational Intranet, for example), the author may be able to make some reliable assumptions about the range of software tools which that user constituency has at their disposal and which they might employ to view that record. The author can not, however, know how individual users may adjust the values of parameters in their separate executions of the same viewer program. The

potential for variations in the values of these parameters means that it is not possible to predict with certainty that the presentation of a document experienced by a reader will correspond exactly to the expectations of the author, and that it is quite possible that two readers of the same representation at the same moment in time will experience differences in its presentational form. A good example is the reader who uses a global command to change the point size which will render the document either longer or shorter than the original and can also result in the removal of logical formatting.

However, although these different 'use contexts' of the same document may generate or employ (or even require) **different** sets of presentational conventions, those conventions are used to convey the **same** logical structure.

In short, it can not be emphasised too strongly that **the presentation of a document is not the same as the structure of the document**. Presentation provides a means of communicating information about logical structure. Indeed it is the logical structure of a document which determines presentation (in conjunction with other factors such as limitations of the rendition tools and the inherent constraints of the target medium, and also perhaps aesthetic considerations). And **it is this logical structure of the document-based record - rather than the specific conventions used to communicate that structure within a single use context - which must be preserved**. This can easily be seen by comparing different editions of some published works - the cheap paperback edition will have the same logical structure as the initial hardback edition but will have a completely different representational form.

An approach to the preservation of the record which accepts the premise that both the physical structure in which it is stored and the presentational form in which it is viewed will vary represents a significant break from the practice and expectations associated with the preservation of the paper record.

It also requires that this logical structure - which is communicated from author to reader only indirectly through context-dependent cues - **is described independently of the changing physical structure and presentational conventions**.

1.5.4.3 Separating structure and presentation

Presentational conventions will vary depending on the medium in which a representation of a document is rendered.

It is also frequently the case that the means by which presentational characteristics are created in a given display medium is dependent on the particular software used to render the document. This is the case because programs such as word processors implement formatting by embedding in the content of the document instructions to format sections of text in a specified way, and those embedded instructions employ a coded form which is specific to that authoring/formatting program (the specification for which is rarely in the public domain). Thus the representations created by such tools are typically limited in two ways: they incorporate presentational conventions appropriate for a single output medium, and they require processing by a single software tool. This can be illustrated by the fact that one word processing package will not read a document created in another without transformation unless it has been saved in a common exchange medium such as .rtf or ASCII where there is no embedded formatting. Another is the way in which MS Word sometimes arbitrarily removes formatting if a block of text is pasted between documents.

These disadvantages of such 'procedural markup'[11] systems were recognised as long ago as the 1960s, in the context not of preservation over time but of the exchange of documents between the processing systems of commercial publishers. The solution they proposed was the use of

[11] Markup is simply text that is added to the data content of a document in order to convey information about it. Procedural markup consists of instructions to an agent (traditionally, a human typesetter, now more usually a computer program) to perform some process on the data, typically to format a piece of text in a certain way.
The 'classic' work in this area is: Coombs, James H., Allen H. Renear and Steven J. DeRose, 'Markup Systems and the Future of Scholarly Text Processing', *Communications of the Association for Computing Machinery* 30/11 (1987). Available at http://www.oasis-open.org/cover/coombs.html

'descriptive' or 'generalised' markup systems that explicitly describe only the logical structural components of the document. The (medium-dependent) rules for the formatting of those logical components are composed and stored separately, as a 'stylesheet', and indeed a typical application may employ several stylesheets to provide different formatting appropriate to different processing environments and different delivery media.

The value of the descriptive markup approach was recognised to the extent that in 1986 it was formalised in an open standard, the *Standard Generalized Markup Language* (SGML), owned by the *International Standards Organisation* (ISO). SGML specifies a generalised scheme for describing the logical structure of documents in a system-independent and platform-independent manner. SGML itself is not a markup language: rather, it provides a formal notation for the definition of markup languages for specific document types. *Extensible Markup Language* (XML), a recommendation of the World Wide Web Consortium, is a subset of SGML designed for ease of processing.

Many discussions of digital preservation recommend representations encoded according to such standards-based forms for long term storage, on the grounds that firstly they are free of the limitations and dependencies inherent in the use of proprietary encoding formats, and secondly they are inherently less volatile/more durable and so 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 standards-based long-term storage format. This range of 'delivery formats' required for user access will vary through time.
- The implications of this separation of structure and presentation - the fact that in even the simplest case, the user's experience of the record becomes the product of a process acting on at least two distinct objects (the 'document' and the ' stylesheet') - must be considered very carefully in the context of the requirements for verifying integrity and authenticity.