National Preservation Office

Changing Trains at Wigan: Digital Preservation and the Future of Scholarship

Dr Seamus Ross, HATII, University of Glasgow

NPO Preservation Guidance
Occasional Papers

Author Dr Seamus Ross, Humanities Advanced Technology and Information Institute (HATII), University of Glasgow
Design The British Library Design Office

November 2000

Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*

1. Introduction

This paper examines the impact of the emerging digital landscape on long term access to material created in digital form and its use for research; it identifies and examines challenges, risks and expectations. It is 6:20 on Monday morning. The Virgin train taking me towards St Helens just lurched from Glasgow Central Station. As the train races southward through what will soon be the mist of Dumfries and then the beautiful rolling countryside of Cumbria, I know I face four uncertainities: will my train arrive on time; will the battery in my laptop last the three and a half hours; more troublingly, how am I going to explain to the Museum I am about to visit that the system designs they have produced will not translate into viable and effective computer-based interactive programmes; and most importantly for you, how am I going to convert the weekend's worth of formal notes, jottings of ill-formed ideas, and unrecorded musings into a coherent examination of digital preservation and its impact on research and learning in the future. All four fears share the same roots: they are about using past experience to manage the uncontrollable, the need to prophecy about the unpredictable, and they are about technology.

As I finished that first paragraph the conductor checked my ticket and advised me that for St Helens I would need to alight at Wigan and catch the 9:47 to Liverpool Lime Street. Suddenly I remembered the thrill of changing trains at Wigan in February a dozen years before when, as a postgraduate, I was darting across Britain to record material in museums and archaeological units for my study of Anglo-Saxon dress pins (1992). 'Thrill of changing trains at Wigan', I hear you chuckle. It is easily explained. My postgraduate career began studying history at the University of Pennsylvania in the early 1980s and as a Teaching Assistant

I led undergraduate seminars in British History. Each weekly seminar took as its starting point a core text. One of the most memorable of these had been Part I of *The Road to Wigan Pier*: for the students all the statistics and summary analyses gained context and meaning as George Orwell brought the bleak world of 1930s working class industrial Britain to life.[1] As J.N.L. Myres, a scholar well known for his studies of Anglo-Saxon pottery, explained in a synthetic study of *The English Settlements*, his own understanding of how the world of the early Anglo-Saxon benefited from walking across their landscape and letting his imagination wander.[2] In a similar fashion, the harsh industrial north Orwell had captured gained new life when I walked through dreary grey Wigan in 1986 in vain search for the Pier.[3] For Myres, it is unfortunate that the imagination, which had helped him understand the Anglo-Saxons, had not been brought to bear on his pottery studies. His decades of examining and drawing pots, resulted in a typological study which had it been completed 50 years earlier would have been an achievement, but by the time he had finished it and it had been published in the early 1980s it was out of touch with the evolving understanding of pottery and the focus of Anglo-Saxon studies. These examples all raise issues of context, purpose, the role of the personal in creating and presenting interpretations of our cultural record, and the impact of the change in prevailing scholarly methods, objectives and interests on scholarship itself.

The preservation and re-use of digital data and information forms both the cornerstone of future economic growth and development, and the foundation for the future of memory. We are increasingly aware of the economic value of information and the variety of ways it can be repackaged, marketed, and re-used. This paper is concerned not with the effect of digital[4] data and information on economic growth, but with the influence it will have on the memories of who we were. Investigations and views of the past depend upon access to information and wherever possible access to contemporary information. In the past these records, created as chronicles, poems, accounts, legal documents, letters, diaries, pamphlets and many other formats, were inscribed in clay, chiselled into stone, and written on papyrus, animal skin, bone, tree bark and paper. Computers and network-based communication, and the technologies that they enable, such as databases, geographical information systems, electronic mail and web-based interactions, are transforming the records we create, how we create them, and how we keep them.

Cultural Transformation and Digital Dependence

This increasing dependence upon digital information is having several dramatic effects. First, it is changing the way in which our culture is recorded. For example, there is no longer necessarily a direct relationship between the way data are physically stored, the logical structure in which they are represented for storage, and their interpretation. Second, our culture itself is being transformed. The internet has created an environment in which new communities and social groups can evolve, as well as protocols and etiquette governing virtual social interaction.[5] Simply, digital information is a cultural product. As we think of physical products of culture as artefacts, so we should also be thinking of digital and electronic products as d-facts (or e-facts). These new products form an essential fragment of our cultural record.[6] D-facts are fragile. They must be interpreted (using software and hardware) before they can be manipulated or rendered for display (or printing) and as a result, in their raw form they are of little value and often meaningless.[7] Information stored in digital form is as delicate as archaeological remains of flora and fauna - it is rare to discover them, the environmental conditions under which they were deposited influences their survival, their recovery and study depends upon substantial investment of labour, and their interpretation requires a vast array of scientific techniques. Preservation of digital information requires active intervention; left unsecured it is susceptible to loss through the physical breakdown of the media, rendered inaccessible by technological advances, or left meaningless through lack of or insufficient contextual evidence.[8]

'Where there is an economic advantage in re-using information or a legal requirement to retain it there will be an easy business case for its' preservation.[9] The short term value of digital information or records often reflects particular regulatory environments or specific types of industry or organisation. In the financial sector records are retained on average for seven years, but in the nuclear and pharmaceutical industries records are required for business purposes over far longer periods. Data created as part of drug development (e.g. clinical trial data) needs to be retained for decades to comply with requirements established by regulatory authorities.[10] Nuclear dump and reactor data will be valuable for hundreds if not thousands of years, where, for instance, it provides evidence of contaminated land.[11] In the automotive and aircraft industries engineering and manufacturing data need to be retained for the life of the product (and not just during the period during which the product is being made) to protect companies in the event of

product liability lawsuits.[12] Satellite images are a good example of unique time sensitive data that cannot be recreated when lost. Where images taken by NASA in the 1970s of the Amazon Basin are still accessible, they contribute evidence to the understanding of change over time in rainforest coverage and density.[13] Even where records can be recreated from their analogue originals, it is expensive to do so. At the 1995 meeting of the ISO Archiving Standards working group it was reported that it cost (including labour) about $5-7 per megabyte per year to retain electronic records created in the engineering sector, but about $1250 to reconstruct them if they were lost or destroyed.[14] Petroleum survey records are even more expensive to recreate. The National Archives of Australia hold 600,000 computer tapes containing oil survey data. These data are regularly re-used by oil exploration companies; recreating the off-shore data would have cost in the early 1990s AUS$10,000 per metre or AUS$10 billion in total.[15] Of course, the possibility of recreating information from analogue sources applies to a decreasing percentage of digital data as much tends, like satellite imagery, to be born digital and never to have existed in analogue form.

Evidence and Memory

It must be obvious that the electronic aspects of our culture targeted for preservation are those materials that have evidential value in the event of litigation, have academic or commercial re-use value (e.g. digital film,[16] academic research data or petroleum survey data), or are valued for their contribution to corporate or national memory (e.g. in the United States presidential emails). Little of the web-based literature[17], net-based advertising, online databases, newsgroups, chat-rooms, virtual communities, music recordings, websites (including webcams), and digital images, which characterise the creations of the several hundred million internet users, are being preserved. This is hardly surprising in such a fluid environment.

Digital networks and the services they support, including for instance electronic mail (email), have transformed communication practices.[18] We can exchange messages faster and communicate with a broader community and with individuals drawn from across much of the globe. The kinds of materials we can use to construct messages now ranges from text and audio to moving images and virtual reality. Just as surviving correspondence enriches the writing of social historians so their digital equivalent, email, will provide a valuable source for future historians.[19]

In reconstructing the past, historians have always used a great diversity of sources. For example, Jean-Jacques Aubert investigated papyri from public and private archives, epigraphic inscriptions, archaeological remains, and literature to produce his elegant study of Roman Business Managers.[20] The dry facts stripped from archival documents were given life through the judicious use of satirical quotations from Roman playwrights (e.g. Plautus). In common with other historians, he pulled together the residue of the past: evidence distilled and teased from fragmentary documentary sources, artefacts, and art historical materials. We recognise that to know who we are and the impact of our past on our future, we need to know where we came from. As a consequence the approaches to constructing stories about the past and the kinds of stories historians build are varied.[21] In our multi-cultural society this richness of approach reflects the resources we are creating and the expectations of the world in which we live. Historians of the future will seek out chat-room transcripts, newsgroups, emails, webcams, and websites along side company and government records, and credit and health data sets.

Both accident and design contribute to ensuring the survival of historic records. The burning of the Persian palace-site of Persepolis after its fall to Alexander, for example, although a savage act of vandalism, contributed to the preservation of the palace archives. In the 1930s excavators recovered this archive inscribed on unfired clay tablets. Under most conditions clay tablets are, by their nature, more durable than other types of media. The Persepolis tablets were written to track economic transactions. The scribes who recorded them would perhaps have been surprised that by analysing these thousands of tablets, it proved feasible to profile the position and role of women in ancient Persia under Darius I to Artaxerxes III.[22] Sadly, only a percentage of these tablets have been fired since their discovery in the 1930s and many are reported to be drying out and crumbling away in their new home at the Oriental Institute at the University of Chicago. The content of many of the tablets has not yet been transcribed and mere recovery of media does not necessarily protect it or its contents against loss. Indeed, recovery may, in fact, expose the material to new dangers. Likewise, we move our digitally encoded information from one decaying medium to another equally unsuitable media type or from one archaic encoding format to another soon to be superseded format.[23] The medium and form of clay tablets, and their content and its structure, raise comparative issues of media durability, content visibility and intelligibility, metadata[24], 'contextual metadata drift', data recovery, and technological

obsolescence.

The structure, form and incisions on clay tablets indicate the presence of information, even if we prove unable to decipher it. Digital storage media, on the other hand, give little clue to the presence or format of the data they contain, and only occasionally an indication of what devices might be needed to access them.[25] Preserving digital assets cannot happen as an after thought, it must be planned: media degrade (e.g. magnetic particles lose their properties and dye layers on optical media break down),[26] technological developments make systems obsolete[27], or information is rendered inaccessible by changes in encoding formats.[28] The short term economic and productivity advantages offered by digital storage, manipulation, and communication encourages us to depend on them more and more. Although some are aware of the preservation risks, society in general is ignorant of them.

2. The Landscape

We have reached a turning point in the production, distribution, and handling of recorded information. Resources are created in digital form:

- that can be easily distributed and refreshed;
- that do not suffer loss of fidelity when copied or used;
- the integrity of which can be secured and verified;
- that can be analysed using an array of automated processing tools; and,
- that can be searched with increasing degrees of sophistication and accuracy.

Contemporary society creates two classes of digital material: the results of retrospective conversions, and resources created only in digital form.

Society and Scholarship

The belief that society and scholarship benefit from the availability of high quality digital resources and networked access to them, is reflected in the increasing emphasis that academic institutions, public archives and libraries, commercial media companies, and funding agencies place on the retroconversion of material.[29] Watching these activities one senses the wholesale rush to retroconvert our documentary heritage into virtual form without a realisation that we are potentially exposing this virtual material to an increased risk of loss. Although branded with

its own inherent risks, digitisation offers an excellent way to address the preservation and access problems posed by material on acidic paper. Millions of archival documents, architectural drawings, newspapers, photographs, film and audio recordings and many other materials are being digitised. Where digitisation is inappropriate project staff are keying in the data. Retroconversion is an expensive process. The resource implications and the associated technical challenges mean that contemporary librarians and archivists face major problems in deciding what of our cultural heritage should be retrospectively converted.[30] Hundreds of retroconversion projects of all different sizes and types are now underway.[31] Hundreds of millions of euros are being invested in these projects. The diversity of approaches, the variety of technologies and management practices, and range of funding models are reflected in these ten example projects: The British Library's Beowulf Project,[32] The Archivo General de Indias,[33] the Music Performance Research Centre,[34] the Cornell Digitisation programme,[35] Beazley Archive,[36] the Scottish Archives Network (SCAN),[37] JSTOR,[38] American Memory Project,[39] Duderstadt Archive,[40] and the Bibliothèque Nationale de France[41]. These and many other projects are increasing the diversity of resources accessible to researchers and the general public. The influence that these projects will have on education, life-long learning, and scholarship will only become clear with hindsight much as happened with the substantial editing enterprises of the nineteenth century (e.g. Rolls Series, *Monumenta Germania Historica*).[42]

Approaches to publication are changing. A shift towards large scale databases, in which the results of the efforts of individual research teams can be aggregated for comparative and often interdependent analysis, has begun to take place in the sciences.[43] The Human Genome project, which is generating a digital record of our genetic makeup, is the flagship of this kind of approach.[44] Another example, the Protein Information Resource (PIR) has, since 1984, provided researchers in evolution and computational biology with access to a regularly updated, quality controlled, and detailed protein sequence database.[45] In the humanities, projects such as the Corpus of Romanesque Sculpture in Britain and Ireland (CRSBI),[46] and English Heritage's Images of England[47], are producing digital image and data banks to improve our understanding of heritage assets and to enable new ways of investigating them. The issues of sustainability, maintenance, and enhancement are critical to the viability of these resources. These databases will be maintained so long as they have immediate public or research value, but they may not continue to

be when immediate interest in them has diminished. Yet we know that information assets go through phases of value and they often regain their value in the future. Unlike many analogue resources that can survive periods of neglect these may not.

Internet Archaeology, an early and continuing experiment in electronic journal publication, reflects a radical rethink of how the results of scholarly endeavour could be presented.[48] While the electronic environment changes how research can be constructed and presented, it exposes this material to risk. Libraries do not keep copies of *Internet Archaeology*: long term access depends upon continuity of the original project. The only copy (save a few backups stored on magnetic tape in the Special Collections of the University of York Library) is the one held by the Project itself. Even though *Internet Archaeology* has rigorous data management policies, this lack of redundancy and geographic spread of copies increases the risk of loss. This problem is exacerbated by the fact that the ability to read many articles depends on specialised software (e.g. use of Cosmo player for the Virtual Reality Modelling Language (VRML)) which the project may not be able to archive.[49]
A lack of strategic planning on the part of electronic journal publishers means that results of research presented in electronic form are at risk of becoming inaccessible. Moreover, as the sources on which we build our research are increasingly digitally -based, the sources themselves will not survive, will be moved to new locations, or will be moved from open access to restricted areas or on to chargeable sites. These changes increase the difficulties associated with verifying conclusions or repeating analyses: the very functionality which digital presentation should enable. One wonders whether the surviving record of scholarship will be like constructing the contents of ancient and medieval libraries from the few surviving manuscripts - scholars can identify the names of the titles of many works - but no unique copy of many of these titles survives anywhere.

Documentary Heritage

Far more alarming than all this has been the change in the way society now creates its documentary heritage. As Rick Barry has noted, changes in working practices are having a substantial impact on the process of document creation, their format, how they are used, and how they are managed.[50] Business activities depend, increasingly, on databases, digital images, geographical information systems, voice mail, email, video recordings, spreadsheets, and word processed documents. As a

result, records (from an archival vantage) and information come in a plethora of types, formats, and structures. The working practices of archivists, librarians and records managers respond to a world composed of textual documents, when the business activities of their organisations now use a range of composite documents of varying complexity:

- static documents composed of such elements as text, tables, and images;
- multimedia or data-rich documents such as the kinds of documents that we encounter in the networked environment (on the world wide web or on www-based corporate intranets); and,
- dynamic documents dependent upon data that might have variable instantiations and be held in databases and spreadsheets.[51]

These materials form the major resources for the future of memory.

With some billion webpages currently accessible (August 2000), including thousands linked to databases, the amount of information accessible to contemporary users of the web can only be described as astonishing.[52] Similarly, millions of email messages are exchanged each day; at the US Department of Health and Human Services at least a million email messages are exchanged each day within this department alone.[53] Financial institutions, commercial users, and governments move even larger amounts of data across private networks and much of this is shipped from computer to computer without any human intervention. Airlines, banks, and credit card companies handle billions of transactions each year. The data contained in the detailed record of each individual transaction provides information that could be used to profile many aspects of modern society, which more traditional textual documents do not illuminate. The sheer quantity of the data, however, makes its retention difficult even in a world of falling storage costs.[54] Where they are retained, these data could be mined in different ways.[55] Resources might be examined by considering change over time and credit transaction data or travel information linked with, say, medical records. Researchers will use the surviving information, not to understand the individual transactions recorded in the data, but the society(ies) that created them. The proven potential of data mining methods indicates that even where data have been summarised for contemporary purposes, the unaggregated data should be retained because future analysts may wish to use the primary data for new purposes. Our scribes in Persopolis retained the records of the individual transactions after they

produced the summary records, yet it was the individual transactions which yielded their riches to scholars by allowing them to use economic data to develop cultural understandings.

From the early 1990s onwards, attention has been drawn to the parochialism that pervades studies of electronic information.[56] Archivists and records managers have focused on the preservation of records, and librarians on ensuring long term access to published resources and online databases. Yet, until we have a much broader acceptance of the realisation that it is essential that we preserve a broader cross-section of our digital heritage, media manufacturers, systems developers, and software designers will not be encouraged to introduce preservation capabilities or functionality into their products. The discussions of the issues of digital preservation must be taken beyond the traditional communities of archivists and librarians if progress is to be made in addressing the dangers.[57]

Environmental Applications

Major new initiatives are collecting data to improve the contemporary understanding of the impact of human activity on the environment and man's changing relationship with it. These resources will provide a foundation for studies of change over time. In monitoring natural phenomena (e.g. seismic or meteorological activity) or experiments, scientific instruments can capture gigabytes of data every hour.[58] As a result, it is not uncommon for data sets as large as a terabyte to be collected and it would not be impossible for a sequence of monitoring activities to lead scientists to accumulate a petabyte of data. In turn the precision of many simulations depends on hundreds of gigabytes of data. In these data sets resides information about bio-diversity, environmental conditions, and our genetic makeup. Large numerical data sets and text files represent a fraction of the uses and products of the digital environment. Designers of buildings, airplanes, and a host of products have increasingly turned to computer environments to enhance the design and testing processes. Architects and mechanical and electrical engineers depend upon computer aided design/drafting applications and virtual reality systems. A consequence of this is that architectural historians will wish to have access to this material. Unlike numerical data which can be migrated from one software/ hardware environment to another, access to this material will be dependent upon access to specific software tools and its interpretation will depend upon the use of original hardware or the emulation of that equipment. This is especially true of

Virtual Reality (VR) models. Looking back, historians will wish to see the materials in their original context, whether this must be the actual environment, or whether a virtual one would provide a suitable context, remains an open question. The answer will depend on the nature of digital experiences in the future.

The use of computer technologies for entertainment and in particular interactive computer games is recognised as a major driving force in the race to create faster computers with better graphics.[59] Only a very small number of institutions are collecting these games and their consoles, yet they are critical cultural artefacts, as they are shaping leisure behaviour and have provided the training ground for a whole generation of computer users. The violence inherent in many of these games may be playing a role in shaping our societal attitudes towards violence; the preservation of the games will contribute to future researchers' understanding of our culture.

Connections and New Technologies

On current evidence, historians of the future will be left with a large number of disconnected d-facts that will prove difficult to use. In the first conference[60] to look at these issues, we investigated such questions as to how much data should be retained: all possible records or just a sample of them. One of the participants, Professor Martin Campbell-Kelly asked, for instance: Would a single airline reservation transaction have value to a future historian? Would it be practical to retain all the tens of millions of airline reservations for posterity? Would future historians be overwhelmed if we were to preserve all these data? Similar questions could be asked about other classes of records. Confronted with the realisation, that faced with the vast quantities of surviving paper documents few researchers can be comprehensive, scholars often suggest that as we increase the quantities of digital documents they will be even less able to investigate the material in a comprehensive way. This is a spurious analogy. Archives of the future will be different and researchers will adopt new, and more technology dependent, ways of working. With an array of analysis tools they will work more exhaustively with the surviving digital resources than they have so far been able to work with analogue ones. Insurance, retailing, and banking sectors currently exploit the potential of data mining techniques and tools to extract information from large heterogeneous data sets.[61] Increasingly, research efforts to interpret these data are assisted by data visualisation tools. In addition to data mining and visualisation

tools, future researchers will be aided by intelligent agents that explore the Internet (or its successor) looking for information that meets certain user-specified criteria and refining their searches as they accumulate data and knowledge.[62] Digital archives combined with new technologies will liberalise scholarship. They will enable simultaneous access to a range of sources (both local and distant) and facilitate the use of research methods not possible with conventionally printed or hand written records.

3. Digital Preservation: A Proactive Approach

This vision of a rich information record just waiting to be harvested and processed by the technology-enabled researcher of the future depends upon the survival of digital data. Sadly, based on current experience, it is evident that not much of this digital material will survive. It is already impossible to find old documentation for early computers, such as those from the 1960s and 1970s, even if you can locate an aging machine.[63] Access to material created using superseded operating systems (e.g. CP/M) or word-processing (e.g.Wordstar) and database (e.g. Dbase III) applications is difficult. Legacy systems written in such languages as Cobol, PL/1, and Fortran are equally prone to loss and were frequently tied to particular operating systems and system calls (i.e. application interfaces to the system level routines) to particular hardware devices.

In comparison to resources produced on analogue media, such as on paper, resources created in digital form are fragile and easily prone to becoming physically and logically inaccessible. The degradation of the media on which they are stored, loss of functionality of access devices, loss of manipulation capabilities, loss of presentation capabilities, or weak links in the documentation chain, are all factors that contribute to making resources inaccessible.[64] Other factors such as loss of contextual information or relevance can render resources non-interpretable. While it is true that it is feasible to recover data that have become inaccessible under many circumstances (and even to the surprise of some data creators as happened with Iran Contra data), this can be an expensive, labour intensive and a risky approach to the resource preservation problem, especially when we recognise that few data recovery methods are comprehensive or successful on every occasion. One risk is that while much data are not 'bit critical,'[65] the same claim can not be

made for most software. On the other hand just ensuring that the binary digits are intact through refreshing the media will not ensure that the digital resources can be retrieved, interpreted, manipulated and presented. Work by the Bundesarchiv (Koblenz) to recover the *Kaderdatenspeicher*, a detailed database of East German Party Officials, depended upon the survival of contextual information, and documentation.[66] In general the digital resource is such a new concept that we have not yet had time to come to grips with its qualities in an effective and meaningful way.

Old Strategies and New Problems

In the past librarians and archivists have worked to ensure that the resources we need are available for very long periods by collecting, documenting, securing, and managing them. If digital materials are to remain accessible over decades, let alone centuries, preservation features need to be incorporated into them. Wherever possible, their preservation must be an integral element of the initial design of systems and projects.[67] This rarely happens. Most digital preservation work, as explained above, must be carried out after the resource has been created and frequently when it is no longer an operational system.

Records managers have continued to raise awareness within companies about the dangers posed to corporate memory by the increased use of digital technologies, if preservation features are not inherent and designed into new systems. The viability of long term retention of digital materials depends, they argue, upon records management involvement in the design of new systems. Current practices do not recognise the value of the participation of preservation specialists in the system design phase, so the effort of records managers focuses on rescue after creation and then generally after acquisition. Even by the time digital materials are passed into the care of records managers, the systems and computer hardware on which they were created are often obsolete.[68] The records arrive on any number of media: tapes (e.g. nine-track tape of varying bit-densities); cartridges (e.g. TK50, DLTs, DAT); hard disks; floppies; solid state storage devices; or CDs. In each case a range of generations of media might be included in the deposit (e.g. 8", 5.25", and 3.52" floppies), magnetic tapes and cartridges. Of all these media, the CD-ROM or CD-R is probably the easiest to handle because at least the standards have been broadly in place for nearly two decades and the drives are ubiquitous, although the low stability of the media may prove an

obstacle. In other cases, the media may prove inaccessible because the peripheral devices are no longer available. How many organisations have access to 8" floppy drives or quarter-inch cartridge (QIC) tape drives and appropriate software, and the software drivers to operate them. Even if it is possible to get the data from the media it will be in a range of file formats including word processing, sound, text, image, and database file formats. Little will be generic enough to be accessed without the original applications. Some of these will be proprietary and others will require particular versions of software that was long since superseded.

Who is selecting and how they are selecting material for preservation is undergoing change. Records managers, archivists, and administrators have long recognised the administrative, legal, and information value of records as well as their enduring research potential as key selection criteria. In the new technology environment, appraisal decisions might come to reflect technical issues including the quality of data set/resource documentation and (or) metadata, and the uniqueness, or rather, the ubiquity of the operating system, software or hardware environment needed to access/use the data.[69] Researchers can only hope that documentation is preserved so that environments can be reconstructed and the integral relationships within the discrete data units necessary to render the information resource processable can be re-established.[70] The problem is not so much that it is impossible to retain all information created in electronic form, but that it is not feasible to document it suitably to ensure its long term accessibility. Unlike the Bisutun Inscription or the Rosetta Stone, where a single inscription helped unlock the records of an entire culture, much more will be required if digital materials are to be accessed in the future.

Indeed part of the problem facing archives and libraries wishing to accession electronic publications is the heterogeneous nature of the materials and the lack of consistency in data, application design, documentation, and metadata. Recent efforts have concentrated on establishing digital preservation infrastructures that are platform independent and support ingesting of heterogeneous digital resources. As the researchers in the NEDLIB project have concluded, solutions must be founded on a layered architecture, 'that provides a clear separation between the hardware, such as storage and communication devices, the protocols and the applications', and depends upon open standards[71]. This invariably means that material will be selected for preservation and that metadata, or information that is attached to primary data to give them context and usability, will have an impact

on this selection process.

In any preservation strategy, metadata (or data about data that makes that other data useful)[72] will have a pivotal role as it provides the only way to capture the context of a resource and the processes defining and surrounding its use.[73] The issues associated with generating metadata sufficient to ensure digital preservation remain unresolved despite the efforts of numerous research projects. The work done on preservation metadata, although extensive, has been primarily at the theoretical level and little of it has been adequately tested on any systems of scale.[74] Metadata is a central element of any model designed to ensure that preserved data is functional. One of the problems, however, is that the complexity of inter-relationships between resources and the various software applications used to run them, may be easily overlooked when creating the metadata elements of the wrapper. The representation of metadata (e.g. Reference Information and Preservation Description Information) also raises difficulties. At its highest level there are three types of metadata: preservation; bibliographic or discovery; and administrative or management.

The creation of software and hardware independent records or 'information assets' requires that all materials that are placed in the archives are linked to information about their structure, context, and use history. These metadata must be sufficient to support the migration of records through various generations of hardware and software, to support the reconstruction of the decision making process[75], to provide audit trails throughout a record's life cycle, to provide records with self-selecting and self-appraising characteristics, and to capture internal documentation. Conceptually, metadata that needs to be attached to or associated with digital information and especially electronic records includes:

- information about the source of the data;
- details of how, why and when it was created;
- details about its intended function or purpose;
- guidelines about how to open and read the record;
- terms of access;
- the migration history of the record and any changes made to it after it was created; and,
- information about how it interrelated to other software and records used by the organisation or other organisations.

Within this emerging consensus there are difficulties:

- metadata models have not been tested across a matrix of organisational categories and data types;
- using metadata to represent the business processes (including information flow);
- associated with layers of documentation and the need for a wider coverage of system documentation;
- incorporating metadata guidelines into software; and,
- associated with ensuring future systems will be able to interpret and use these metadata.

The current definitions of the form and properties of the metadata that are useful for the management of other data, often ignores issues of process and the need to define metadata in terms that reflect the specific environment in which the data was used. Metadata need to be derived from an analysis of the organisational functional requirements and needs. Only where metadata categories are derived from an analysis of the business processes, can one ensure that they reflect the functional uses made of the data. Concentration on the definition of metadata divorced from the processes that need to be undertaken using metadata, will result in the creation of metadata guidelines of limited value because they will not reflect the data environment. Where the metadata include evidence of the processes, future users will be in a strong position to understand the role that information/ data played in the organisation. The process driven model underlies the research being conducted by the DLM-Forum Working Party on functional requirements.[76]

Records exist for a purpose and these reflect the functions and activities of the creating organisation. Appraisal and selection criteria include the evaluation of the relationship between the purpose for which the record was created and the organisation's purpose. Purpose is closely related to process. Process descriptions are essential if records are to be worth retaining. This is true not only for records preserved as a record of the creating organisation's own history, but also for records retained for their informational value, because they need to be contextualised in the environment in which they were used. Appraisal and selection of electronic records will also involve an evaluation of the processes that were performed on the metadata, to reconstruct the original organisational processes and behaviour.

We now widely recognise that preservation is an active process and the debates

about digital preservation strategies have generated a substantial quantity of literature.[77] There are numerous preservation models emerging and even systems under development, some in the libraries arena,[78] some within the sciences and social sciences,[79] and still more in the area of archives and records management.[80] The main objective of all this effort is to ensure long term access to digital resources created by contemporary commercial companies (e.g. publishers) and public sector bodies. We expect to be able to identify, retrieve, render, and use these resources in the future. Further, we expect those resources to be complete, authentic, and verifiable.[81] Preservation strategies that have been investigated include preservation of obsolete technologies,[82] migration of digital records to new environments,[83] emulation of obsolete systems (e.g. applications, software, and hardware),[84] bundling,[85] persistent object preservation,[86] and binary retargetable code.[87] Almost without exception, preservation strategies depend upon refreshing media through replication or copying of information to new media types.[88] None of the currently available preservation solutions could be described as 'tried and tested', but work in developing modelling tools, models, and standards have begun to provide a more secure foundation for ensuring the longevity of digital objects.[89]

Relying on a single preservation strategy is analogous to single crop economies in an agrarian society. For instance, migration has been suggested as a primary method of digital preservation and it has been more widely reviewed than any other approach.[90] It depends upon the periodic conversion of resources from older forms into newer formats before the older formats become obsolete.[91] The appearance and behaviour of digital resources is frequently intrinsically linked to proprietary software and even hardware. For all the discussion about migration the small scale of the test-beds, has not yet provided sufficient evidence to establish with any degree of confidence, that such a model could be applied across a heterogeneous and changing resource base. It is this variability in the resource base and the rapid changes in the digital models that make relying on migration strategies alone risky. Furthermore, as migration is very much a handcraft, it makes the process of migration-based digital preservation labour intensive and therefore expensive. These costs will increase the greater the time depth between the point at which the digital asset was created and when it is to be migrated.[92] The size of the resource does not necessarily correspond directly with the cost of its migration or emulation. A small resource that required significant labour investment could cost as much to migrate as one hundreds or thousands of times larger, but for

which less intervention was necessary.

Often overlooked are the twin issues of one, acceptable levels of loss and two, the complexity of designing and implementing automated testing strategies to ensure the functional relationship between the digital materials before and after migration. Before we can see migration as a viable aid to preservation, more work is needed in the development of metrics for benchmarking and supporting the evaluation of the risks to the functionality of the data set, or losses resulting from particular changes. The question of 'how much loss is acceptable', whether this is in functionality, integrity, authenticity, or meaning has not been adequately addressed by any commercial or research initiatives.[93] Part of the reason is that conceptually, any loss is not acceptable, but practically, it is difficult to operate preservation methods that are entirely loss-free. If the primary aim of digital preservation is to ensure that future users can identify, retrieve, render, manipulate, and use resources then it is essential that loss of content, context, integrity or functionality be negligible or at the very least quantifiable.[94]

Whatever the longer term preservation methods adopted for an individual resource, all resources will need to be wrapped for preservation. Wrapping will involve encapsulating or linking the resource to adequate reference (e.g. description of data types, operations, relationships) and preservation description (e.g. reference, provenance, context, and fixity) information. The precise metadata requirements of each digital object will vary, and the metadata required for each digital resource could be drawn from a metadata repository.[95] For instance, it is worth recognising that if a digital resource is destined for migration, a different metadata set would be required from that necessary for emulation. Where intervention-based preservation is feasible preservation of digital resources depends upon a mixture of strategies: the OAIS (Open Archival Information System) model supports this diversity of approach and is currently emerging as the standard methodology.[96]

What we are seeking is a strategy for ensuring the perpetual access to digital resources that protects the integrity, functionality, and meaning of digital materials. This can only occur where acquisition and management of digital resources is controlled, contextual information is secured, and sufficient preservation metadata are attached to the resource to ensure it is capable of interpretation in the future. While these are primarily technical requirements, the effectiveness with which these can be addressed depends on the organisational and managerial environment

in which they are to be conducted. In other words, preservation strategies without policies will not work.

4 Digital Preservation: Accident and Rescue

It should be obvious that we actually understand a tremendous amount about the problems associated with digital preservation and have many ideas about how to overcome them. Apathy, lack of realisation of the urgency of the digital data problem, and a shortage of skills is leading to inaction. This trend, which is also exacerbated by the fact that the costs associated with acquiring and preserving digital data are in addition to, rather than a substitute for, the costs of acquiring and preserving analogue materials, will continue. Unfortunately, unless we have urgent concerted action, neither strategically planned nor intervention-based preservation will be the norm; the most common preservation method for digital data and records will be 'accident'. Where digital data do survive, future users will not access them in the same way contemporary users do.[97] Valuable cultural data contained in the record structures, software (e.g. applications), and hardware will hold keys to understanding the material itself, processes of work, and the culture which created the materials. Much as, from close analysis of the formulae, the internal structure and the vocabulary of the *Donatio Constantio,* the Renaissance thinker Lorenzo Valla (1440) demonstrated it was a forgery,[98] future scholars will attempt to validate their digital resources. Scholars will ask, whether severing the data from the environment in which it was created and used debases the message. Even contemporary users recognise that this is the case with multimedia, geographical information systems, databases and records. Without the original software to process and render the data, it will be impossible to determine what kinds of (or specific) records users might have created and how they were presented to them. Furthermore, just as modern historians study ancient archives to understand working practices and the functioning of government administration,[99] cultural and social historians will look beyond the content of the records to understand how the tools and the environment of work conditioned the social behaviour of the workers and work.[100]

From among the long term access methods, emulation is the most promising.[101] It will make it possible for us to experiment with historic hardware and software

and understand old ways of working. One wonders how we will comprehend the misery caused to a generation of users by flickering and immobile screens, and poorly designed keyboards, if we are unable to experiment with the older equipment itself. As far as hardware is concerned, experiments have demonstrated that simulation does not provide the user with the same level of understanding which can be provided by access to the original environment. For example, the Computer Conservation Society has developed a simulation of a late-1950s Ferranti Pegasus Computer that creates a 'Pegasus virtual processing environment' on PC-architectured computers. While the simulation enables users to perform data processing tasks in a Pegasus-like manner, it does not provide the same impression that the experience of watching technicians run an original forty-year-old Ferranti Pegasus Computer in the Science Museum (London) can give of the process of work in the late 1950s. Other well known emulations and simulations including the ENIAC-on-a-Chip,[102] PDP-11,[103] and the EDSAC[104] demonstrate that we will be able to re-create unique systems.

Emulation offers a secure way of ensuring access to data, software, and presenting the functions of older systems. The research team at HATII, has examined the growing number of sites devoted to emulation and conducted experiments and evaluations to determine the viability of system emulation and simulation. Our small-scale pilots lead us to conclude that emulation offers a viable path. Larger scale experiments are necessary to confirm this assessment. These must involve, for instance, larger data sets, a more diverse range of input devices, more programming languages, programs using a greater number of system calls, applications depending upon the presence of particular hardware, and more complex software. The Digital Repository Project initiated by Hans Hofman in the Netherlands is running a number of scenarios to examine in a strategic way the viability of emulation.[105] Other comparative experiments are needed and eventually the process of conducting an emulation needs to be described in detail.

Even if we do nothing it is unlikely that everything will become lost, although the information value of what might survive will be *ad hoc* and less rich than if positive action had been taken. For example, even after storage under extreme conditions it is still possible to recover data from magnetic media. A team of scientists at IBM demonstrated this when NASA presented it with computer tapes from the wreck of the Challenger Space Shuttle. These tapes had spent six weeks immersed in the sea off the Florida coast.[106] When recovered, the tapes were

covered in salt deposits, the magnetic coating had disappeared in some reel segments and in others the tape substrate was eroded as a result of chemical reactions. The main damage this corrosion had caused was to the adhesive of the tapes and as a result the binder and substrata were separating. Despite these deficiencies the team was able to recover the vast majority of the data after washing and chemically treating the tapes. The project team had several advantages in their favour: knowledge that there was data on the tapes to be recovered; a collection of identical undamaged media to provide chemically controlled samples; the type of hardware necessary to read the tapes; a knowledge of the data format of the tapes; and sufficient financial resources at their disposal. Had they found the tapes, but had none of this information, their task would have been even more difficult, if they had begun at all.

What will certainly be lost where data is preserved by accident is context and we will pass onto the future a major data recovery and reconstruction problem.[107] Techniques of digital archaeology are evolving already. The growing data recovery and computer forensic industries are building tools and methods to find, recover, and interpret data that are inaccessible. Data recovery companies have begun to stockpile components (e.g. hard drive assemblies and hard drives themselves). There is a need for generic devices to read media independently of the read-write devices used to create it. One approach that shows promise combines Magnetic Force Microscopy (MFM)[108] and cryptographic techniques. With MFM it is possible to read the magnetic media. A raw bit stream of 0s and 1s provides few clues to how it is to be interpreted, let alone its meaning, and low level context information such as block size, encoding standards, and file structure, will prove essential. It might be feasible to draw out a text document or a data set relatively easily, but a digital image whose format was completely unknown and appearing as a single bit stream, would require significant analysis before it could be rendered. Here a great deal more research into cryptographic analysis will be required. The nature of digital representation for storage makes data stored in this way particularly susceptible to cryptographic methods of analysis. Of course without associated metadata and other documentation, the full informational value of these recovered data may remain untapped.

New Environments, New Challenges

The internet and the web pose new problems to future historians. Until recently,

most digital information was created inside companies and government organisations, but this is changing rapidly. The internet has provided us with a new and incredibly large environment for information creation and exchange. Preserving web pages and online databases will be challenging, especially where the information that users view is dynamically generated in response to particular queries or information provided. Government organisations have begun to address the preservation of their web pages especially where the web pages are designated as official records.[109]

Yet, the internet is more than web pages and it is its other aspects, such as the social environments it fosters and the communication interfaces on which it depends, that will prove the most difficult to preserve and for future researchers to reconstruct. More dramatically, the internet has enabled an environment in which we can have new kinds of social experiences, where we can take part in new communities, and where the concepts, function and role of imagination, gender, ethnicity, identity and community are taking on new meanings and contexts. Preserving the computer boxes, screens, routers, wires, programmes and applications will not make it possible for future historians to comprehend this phenomena and its transformative impact. The relationship between the function of these objects and the behaviour of the user in general or within specific virtual environments is likely to remain opaque. For example, how will future scholars understand virtual and real world identities, when the former are likely to be multiple, dynamic and unrecorded. This virtual world is changing so fast, that behaviours that were evident and observable five years ago, have disappeared because the environment has shifted the behavioural goal posts. The internet and the culture experiences it provides both reflects real communities and permit us to experience worlds not otherwise open to us.

Other factors are putting digital memory at risk. Fear of legal action encourages organisations to destroy data in their care.[110] Even in countries where the necessity to preserve digital materials is recognised, legal constraints, such as the framing of the EU Data Protection Directive, are making it increasingly difficult to retain certain kinds of data. A significant obstacle to digital access and preservation is inherent in the legal and moral acts taken to protect the Intellectual Property Rights (IPRs) of creators. This dichotomy sets at cross purposes the preservation obligations of today's caretakers on the one hand, and the equally compelling responsibility to respect the rights of data owners. Moreover, the trend towards

ownership of knowledge as though it were a tangible property poses new problems.[111] We are fast becoming a knowledge economy in which data, information and knowledge about how these resources can be used is more important than tangible products. Ownership of knowledge restricts its dissemination and decreases the likelihood of its survival.[112] Alongside these are other risks including encryption and viruses.[113] Encryption is particularly a concern because while we may secure raw data for the future we may make it inaccessible as the decryption algorithms, software, or keys become separated from the data itself. Computer viruses and related programmes, while they may pose a threat to present and future data and systems, need to be preserved responsibly because they tell us much about contemporary counter culture.

5. Conclusion

Increasingly, our culture and its by-products are represented as binary digits. These digital materials are at risk of loss or of becoming inaccessible unless they are properly monitored, managed, and secured. We know what can be done to improve the chances that digital data will survive and the areas where further research could make us even better able to preserve digital materials. It is unlikely, however, that much material will be preserved by design. Accident is more likely to provide the mode of preservation as it has in the past. Whether the material survives by accident or design the survival of documentation about our culture is likely to become increasingly important to a larger and larger segment of the population. In post-industrial countries a growing segment of the population pursues activities that include the study of family history and takes part in a wide range of other life-long learning activities. Their interests extend well beyond merely tracing lineage. They wish to know more about how they and their ancestors fit into the culture, community, and world in which they live(d).

Configuring our Social History

As a consequence, research and scholarship in the future is likely to be more personal. It will be more about configuring stories that tell us who we are and where we came from. It will be about discovering sources and paths to data that help us understand ourselves more fully. Constructing our social history will

include our ability to chart our place in demographic change, to review our medical history and that of our community over time, and even to study the diet of our immediate and distant ancestors. In many ways, history is an attempt to build an interpretative structure on the debris left by others. Over the past 150 years the richness and nature of that layer has changed dramatically, as new disciplines have been created (from social history to demography to women's studies). All this has depended upon new interpretative methods, been related to the changing attitudes of the contemporary world, and most importantly been built on larger and more diverse data, whether these data are artefacts, ecofacts, charters, diaries or a host of other info-forms. The interpretative layers have tended towards the general. New ways of working and new technologies will change all this. Currently through our two classes of digital materials - retroconversion, and new digital content (much of which are the by-products of contemporary life) - we have just started creating the digital record we need for the future. In the not too distant future, we will be able to choose from a virtual catalogue a 'virtual researcher' who (and actually I should have written which, but I have already begun to anthropomorphise our 'virtual scholar') will be able to traverse this digital landscape and collect and analyse all this data/information/records to construct interpretative layers within defined structures. These tireless, nimble, and adaptive knowledge builders will be able to examine massive data sets to study among other information resources, environmental records, chemical dumping and nuclear waste data to help understand better how we have damaged our planet and to help plan its future use. Most importantly, knowledge builders will allow future generations to examine the relationship between their individual ancestors, events and a wider range of societal phenomena.

Ensuring Survival

Any future scholarship depends upon the survival of the digital resources in accessible and intelligible contexts. This is currently a far greater challenge than any study of the past has ever been. How should we tackle this problem? We must engage the popular imagination in the possibilities opened through the preservation of digital assets. We must encourage media, software, and hardware developers to think long term. We must encourage those creating data to recognise their long term value and act to secure them for the future. We must recognise that no single community will address the problems of digital

preservation in isolation. This examination of digital information and its long term accessibility allows us to reach six general conclusions:

- we already know a tremendous amount about how we can assist the longer term preservation of the digital products of our culture and we need to use this knowledge effectively and strategically to improve the chances that they will survive;
- while we can assist future researchers wishing to access data created by our culture by ensuring that our digital products are documented, rich in metadata, and created using open standards, digital information will survive more often by accident than by design;
- during the next decades there will be a growth in digital archaeology and data recovery tools and methods in response to the increasing amounts of unsecured digital information;
- the internet is fast changing all assumptions about digital preservation as it is creating new environments that lead to the creation of significant cultural data outside organisations that are likely to preserve them and it fosters behaviours and interactions that often leave no sustainable traces;
- we need to consider the time-frame for which we are hoping to sustain our digitally encoded memories and decide what is really viable and what is really necessary; and,
- in the short-term, say the next fifty to one hundred years, technological and methodological developments will enable researchers to use surviving digital resources far more comprehensively than has been possible with analogue resources in the past.

Our digitally encoded memories are fast becoming obsolete. As we create more and more ingenious ways of encoding and storing them we tend to exacerbate the preservation problem. While we should act positively to address these difficulties we must not underestimate the lengths to which future generations will go to unravel the record we leave behind. We must also not forget how evocative a single record can be to the future. In 1988 I visited the archaeological site of Boxgrove in Southern England. Some 300,000 years earlier a hunter-gather had squatted below the flint-clay cliffs, which later sealed the site, and made a flint tool. Not long after the fragments of flint, which had been chipped away in the process, were covered by sand. As a result of painstaking effort, archaeologists were able, from the distribution of the flint fragments, to reconstruct the tool that had occupied this ancestor for

twenty or thirty minutes and to even detail which of the tool-maker's hands had been predominant. If we would wish future generations to have a reasonably representative and rich record of our culture, we should act responsibly and collaboratively to ensure that we are leaving a digital record that is durable, processable, and intelligible.

* This paper was originally delivered as the Keynote Address at the Warwick II Conference on 3 March 1999 (Warwick Conference Centre). The original version is available at: http://www.leeds.ac.uk/cedars/OTHER/SRoss.htm.

Acknowledgements: I wish to thank the following for comments. Although I may regret that I did not accept all their suggestions, I wish to thank Neil Beagrie, Nancy Elkington, Vanessa Marshall, Andrew Prescott, and Helen Tibbo for their comments on this paper. I am also very grateful to Belinda Sanderson of the NPO for handling the logistics of bringing this paper to print.

1 George Orwell, *The Road to Wigan Pier*. (London: Secker & Warburg, 1937 [rpt 1980}).

2 J.N.L. Myres, *The English Settlements*, (Oxford: Clarendon Press, 1986), xxv-xxviii.

3 Orwell, 1937, 75.

4 In this paper the terms digital and electronic are used interchangeably when referring to data, information, and knowledge. The term 'records' is used primarily in an archives and records management sense.

5 The Humanities Advanced Technology and Information Institute at the University of Glasgow 1999/2000 seminar, *Investigating Cyberspace: Communities and Cultures on the Net*, examined the growth and nature of communities in cyberspace and the evolution of social behaviour within them. While a number of authors have examined the formation of net-based cultures and have described methods for examining and describing them, these methods and theories remain in a formative stage. (See for instance: Mark Dery, *Escape Velocity: Cyberculture at the End of the Century*, (New York: Grove/Atlantic, 1996); Steve Jones, *Virtual culture: identity and communication in cybersociety*, (London, Sage Publications, 1997); David Porter, *Internet Culture,* (London: Routledge, 1997); Howard Rheingold, *The Virtual Community: Homesteading on the Electronic Frontier*, (London, 1995); and Sherry Turkle, *Life on the Screen.* (New York, Simon & Schuster, 1997)). The more significant aspect of the problem is, however, that as the residue of this culture is digital, its survival is unlikely. There is little that will

be left to future historians from it and even contemporary conclusions are difficult to test and confirm because the pace of change is so rapid and the record so limited.

6 Seamus Ross, 'Historians, Machine-Readable Information and the Past's Future', in Seamus Ross and Edward Higgs (eds), *Electronic Information Resources and Historians: European Perspectives*, (St Katharinen, 1993), 1.

7 While it is true that the content held on materials such as gramophone records and wax cylinders can only be accessed using specialised devices, engineering these devices and processing the data content for output shares none of the complexity of accessing digital data and rendering it.

8 In 1986 the Committee on Preservation of Historical Records released *Preservation of Historical Records*, (Washington D.C.: National Academy Press) see esp. pages 61-69. It was one of the first comprehensive attempts to address the problems associated with the fragile character of information in electronic form. The focus was mainly on media, but the authors did recognise that 'hardware will become obsolete within a couple of decades.' Optimistic promises crept into the report: it was estimated that polyethylene terephthalate (PET) film would last 1000 years.

9 Ross, 1993, 11. A similar argument could be made for digital film and audio because of their market penetration and the proven commercial value of historic recordings and the enduring interest in films.

10 Philip Lord, 'Strategies and tactics for managing electronic data records: a view from the pharmaceutical industry', *INSAR (Supplement II), Proceedings of the DLM-Forum on electronic records*, (1997), 168-174.

11 Nuclear Information and Records Management Association (NIRMA) (http://www.nirma.org/newhome/).

12 A. Herbst and B. Malle, 'Electronic Archiving in the light of Product Liability', *KnowRight '95,* (Vienna: Oldenbourg Verlag, 1995), 155-160. The Boeing 777 was designed entirely on computer. (http://www.boeing.com/news/releases/1995/news.release.950614-a.html). The design data will be required in processable form for the life of the product. In the event of claims of design negligence, it may be essential to be able to use the original processing software and hardware (or emulation) to verify that with the technology available to the designers, they could not have seen phenomena that could be observed when newer tools are used to study the same design data.

13 The National Research Council, in *Preserving Scientific Data on Our Physical*

Universe: A New Strategy for Archiving the Nation's Scientific Information Resources, (Washington D.C.: National Academy Press, 1995, 31) pointed out that much of the early Landsat imagery needed to be rescued before it could be made accessible, but that the work was carried out because 'retrospective data are vital to understanding long term changes in natural phenomena.' The power of time-series data of this kind in helping us to understand the impact we are having on our planet is exemplified by the case study of Rondônia (Brazil). The deforestation can be seen by comparing images taken in 1975, 1986, and 1992 (See http://edcwww.cr.usgs.gov/earthshots/slow/Rondonia/ Rondonia). Another good case study is of the change in the Aral Sea between 1964 and 1997: (http://edcwww.cr.usgs.gov/earthshots/slow/Aral/Aral).

14 (http://ssdoo.gsfc.nasa.gov/nost/isoas/us01/minutes.html). There are other cost models that suggest that annual storage costs are higher, but still lower than the costs associated with the re-creation of the resources. See for example, Charles M Dollar, *Authentic Electronic Records: Strategies for Long Term Access*, (Chicago, 1999), 207-213. Modelling the costs of preservation of digital data remains an area where much more research needs to be carried out. See for instance National Preservation Office. *Digital Culture: Maximising the Nation's Investment: A Synthesis of JSC/NPO Studies on the Preservation of Electronic Materials*, Mary Feeney (ed), (London: The National Preservation Office, The British Library, 1999), 50-60. (Kevin Ashley)

15 See Steve Stuckey, 'The Good Oil for Australia', Barbara Reed and David Roberts (eds), *Appraising Computer-based Records*, Dickson, ACT: Australian Council of Archives and Australian Society of Archivists Incorporated, 1991, pp. 95-104.

16 P.D. Lubell, 'A coming attraction D-Cinema', *Spectrum*, 37.3 (March 2000), 72-78. The quantities of data created through the digitisation or digital production of film are massive: to store the digital version of Star Wars Episode I required nineteen 18 GB hard drives and Toy Story covered more than 300GB of storage. The value though of the preservation of this material for long term access already has a proven business case.

17 While much of the large genera of web-based fiction that is emerging may be of limited merit, it will still have an impact on the growth of net-books and web-fiction and as such may merit preservation. This is the experimental phase from which new genera will spring.

18 J.R. Baron, 'E-mail Litigation Wars: the U.S. National Archivist Strikes Back', Luigi

Sarno (ed), *Authentic Records in the Electronic Age: Proceedings of an International Symposium*, Vancouver: InterPARES Project & Istituto Italiano di Cultura Vancouver, 2000, 156-167. T. K. Bikson and S.A. Law, 'Electronic Mail Use at the World Bank: Messages from Users', *The Information Society* 9.2 (1993), 89-124; Jean Samuels, 'Electronic Mail: Information Exchange or Information Loss?', in Edward Higgs (ed.), *Historians and Electronic Artefacts*, (Oxford: Oxford University Press, 1998), 101-119; David A. Wallace, 'Record-keeping and Electronic Mail Policy: The State of Thought and the State of the Practice', (1998), (http://www.rbarry.com/dwallace.html) ARMA International Standards Committee E-mail Task Force, *Guideline for Managing E-mail* (ARMA, 2000).

19 In the case of Armstrong v. Executive Office of the President, Judge Richey ruled that email in digital form contained information (e.g. transmission and receipt data and links between the messages) that was not present in the printouts. Printed versions were not faithful to the original and were, therefore, no substitute for them (http://www.eff.org/pub/Legal/Cases/Armstrong_v_President/). The principle is that information in electronic form contains details that when it is preserved in any other way than digitally, become lost. It is worth remembering that even in contemporary contexts judges, jurors and lawyers like electronic mail because they believe that the contents of email reflect more accurately the true feelings of the author. A view enhanced by the generally informal character of email. J. C. Sipior and B. T. Ward in 'The dark side of employee email', (*Communications of the ACM*, 42.7 (1999), 88-95) describe the dangers of unrestrained access to email can pose to institutions.

20 Jean-Jacques Aubert, *Business managers in ancient Rome: a social and economic study of Institores, 200 B.C.-A.D. 250* (Leiden: E.J. Brill, 1994).

21 Three examples come to mind: Le Roy Ladurie's *Montaillou* an ethnohistorical study of this Cathar village in South Western France based on the one surviving volume of the inquisition record for the village (1978); Jan Vansina's studies of African history, which were built on oral histories recorded in African communities, and, Nancy Farriss' *Maya Society Under Colonial Rule: The Collective Enterprise of Survival* which investigates culture and cultural change (1984). The work of many other historians (e.g. Trevelyan, Marc Block, Thompson, Thomas, Duby, Braudel, and Le Goff) shows the diversity of approaches historians take when investigating the past.

22 Maria Brosius, *Women in Ancient Persia*, (Oxford: Oxford University Press, 1996). The

work altered our perception of women in the ancient near east: between 549 and 333 BC there is unequivocal evidence that some owned extensive estates and presided over great wealth, some were leaders of large workgroups, and they were remunerated equally with men for the same work. These records made it feasible to demonstrate the inadequacies of Herodotus' account.

23 The Alberta Hail Project for instance transferred gigabytes of data from tape to CD-R, a medium that in many of its types is widely recognised as unstable, see B. Kochtubajda, C. Humphrey and M. Johnson, 'Data Rescue: experiences from the Alberta Hail Project' paper presented at the 21st Annual Conference of the International Association for Social Science Information Service and Technology *IASSIST* 95 May 9-12, 1995, Quebec City. (http://datalib.library.ualberta.ca/AHPArchive/Archive.html). On CDs see Seamus Ross and Ann Gow, *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*, (London, 1999), 11-13; J.S. Kim, T.Y Nam, and Y.J Huh, 'The Optical Characteristics in the Layers of Compact Disc-Recordable', *Korean Journal of Chemical Engineering*, 14.2 (1997), 88-92; C. Södergård, J. Martovaara, and J. Virtanen, *Research on the life expectancy of the CD-R (CD-R levyjen säilytyskestävyyden tutkiminen (Undersökning av CD-R skivors beständighet)* (Helsinki, 1995); but see Henk van Housten and Wouter Leibbrandt, 'Phase Change Recording', Communications of the AGM, November 2000, 64-71. CD-Rs are susceptible to damage through exposure to light, heat, and dampness. Those CD-Rs with data receiving layers made with pthalocyanine rather than cyanine dyes are the more stable. The latter is an organic dye and the former is a metallic stabilised dye. CD-Rs tend to be more stable than CD-RWs. DVD (Digital Versatile Disk) and DVD-R have similar stability issues associated with them.

24 Metadata are data that makes other data meaningful and usable.

25 The carrier (e.g. cartridge, tape, or diskette) does contain markings which may provide an indication as to the media class, the device needed to read it and in turn the hardware and software on which its contents might be made accessible.

26 The literature examining wear, decay through hydrolysis, loss of lubrication, the interaction of the chemicals in magnetic media and loss of magnetic properties is extensive. From among the numerous reports, the following four indicate the scale and diversity of the research into magnetic media: C. Kajdas and B. Bhusham, 'Mechanism of interaction and degradation of perfluoro-polyethers with a DLC coating in thin-film magnetic rigid disks: A critical review', *Journal of Information*

Storage and Processing Systems, 1.4 (1999), 303-320; Y Nishida, M Kikkawa, and H. Kondo, 'Behavior of lubricant migration in particulate magnetic recording media', *IEEE Transactions on Magnetics*, 35.5 (1999), 2451-2453; M.S. Hempstock, M.A.Wild, J.L. Sullivan, and P.I. Mayo, 'A study of the durability of flexible magnetic media in a linear tape system', *Tribology*, 31.8 (1998), 435-441; C. Gao, 'Corrosion evaluation of cobalt based magnetic films using various techniques', *Materials Research Innovations*, 1.4 (1998), 238-242. Some of these developments are very positive: early magnetic media (e.g. 1960-85) had low coercivity and as a result were susceptible to data loss from stray magnetic fields (e.g. magnets in motors such as those that drive the belts in airport security x-ray machines). For example, the coercivity of 720K 3.5" floppies was about 300 Oe, 1.44K 3.5" floppies 700 Oe, and Quarter-inch cartridge (QIC) tape (say the DC600A) rated 550Oe. Newer media has comparatively high coercivity and is as a result less susceptible to stray magnetism. Most media is now higher than 1000 Oe; hard-disc drives made during the 1990s had coercivities ranging between 1400 Oe and 2200 Oe. In the case of tapes, print-through, where the data from one layer is imprinted on the adjacent tape, remains a problem. Signal decay in magnetic media is much less of a problem than is the breakdown of the media binder itself.

27 Newer systems and improved storage devices (e.g. tape drives, discs) lead businesses to replace peripheral devices often without copying older material that is no longer in current use to the new devices. Without these older devices the media often proves unreadable when it is required for long term preservation or evidential purposes. See the Virtual Museum of Computing (http://www.museums.reading.ac.uk/vmoc/), the Computer Conservation Society (http://www.cs.man.ac.uk/CCS/), or Computer History (http://ei.cs.vt.edu/~history/machines.html) for examples of efforts to record maintain, and preserve computer technology. The accumulation of information about obsolete computers and the machines themselves has become a popular activity and the web is littered with sites describing growing public and private collections. See for instance (http://www.obsoletecomputermuseum.org/) and from here it is possible to follow a web ring to other sources.

28 Ross and Gow, 1999, 11-13. The layering of encoding plays a role in interpreting digital information. At the highest level in the hierarchy there is the file of a particular format and at the lowest there are the data as stored on the media itself. Data are encoded as magnetic domains on tapes and discs or as pits on optical

media (CDs). The magnetic domains represent the 1s and 0s as written to the media by the write-heads. These 1s and 0s as stored on the raw media do not necessarily share a one-to-one correspondence with the contents of a particular file. For instance before the data was passed to the controller for writing it might have been compressed (say using LZW [Lempel-Ziv-Welch]) and then the controller itself might write other data before or after the bit stream to assist it to locate and track the data across the media (say in the case of tapes block headers). There are a number of ways of encoding data on to the surface of the media. For instance, Non-Return to Zero (NRZ) is a simple way of recording data on to the magnetic surface. This represents 1 (one) bit by a change in magnetic polarity and a 0 (zero) by no change. Identifying the bit stream and retrieving the bit stream would only be the beginning of the problem. It is then necessary to determine what is encoded. Even for the encoding of text there were competing standards at one time, ASCII (American Standard Code for Information Interchange) and IBM's EBCDIC (Extended Binary Coded Decimal Information Code); and if the sequence of bits turned out to be a character and not a segment of an image it could be 7-bit or 8-bit ASCII or 8-bit EBCDIC. This is just the start.

29 In the United States the Andrew W. Mellon Foundation and National Science Foundation Digital Libraries Programme and in the United Kingdom the Heritage Lottery Fund and the Joint Information Systems Committee are examples of active players. An increasing number of commercial firms, recognising the business potential of owning and delivering digital content, are developing digital stockpiles, including Corbis and Getty Images (see for instance, 'Blood and Oil', *The Economist*, 4 March 2000, 97). These resources enable new research, for example, F.R. Shapiro, 'A study in computer-assisted lexicology: Evidence on the emergence of hopefully as a sentence adverb from the JSTOR journal archive and other electronic resources', *American Speech*, 73.3 (1998), 279-296.

30 Seamus Ross, 'Strategies for selecting resources for digitisation: source-orientated, user-driven, asset-aware model (SOUDAAM)', Terry Coppock (ed), *Making Information Available in Digital Format: Perspectives from Practitioners*, (Edinburgh: The Stationery Office, 1999), 5-27. Elsewhere I have argued that we need national strategies (*Funding Information and Communications Technology in the Heritage Sector*, Policy Recommendations to the Heritage Lottery Fund (January, 1998) (see http://www.arts.gla.ac.uk/ HATII/HLFICT) or in the case of the European Union, trans-country strategies to ensure that we invest in creating unique, interoperable,

and consistent quality digital resources.

31 See the HATII website for a list of these projects at 1 October 2000. (http://www.hatii.arts.gla.ac.uk/projects/digiprojects2000).

32 The Beowulf Project has created a 'diplomatic edition' with transcription, translation and a wide range of additional resources to make this manuscript accessible to students and scholars in a way that will make it possible for them to develop a fuller understanding of its creation, change over-time, and the tradition to which it belonged. See: K. Kiernan, `Digital Preservation, Restoration, and Dissemination of Medieval Manuscripts' in Ann Oakerson, (ed) *Scholarly Publishing on the Electronic Networks*, Proceedings of the Third Symposium, (Washington D.C.: ARL Publications, 1994) 15 (http://www.bl.uk/ diglib/beowulf or http://www.uky.edu/~kiernan/ welcome.html); Andrew Prescott, 'The Electronic Beowulf and Digital Restoration', *Literary and Linguistic Computing*, 12.3 (1997), 185-195; A. Prescott 'Constructing Electronic Beowulf' in L. Carpenter, S. Shaw, and A. Prescott (eds) *Towards the Digital Library: The British Library's 'Initiatives for Access' Programme* (London: The British Library, 1998), 30-49.

33 Pedro González**,** *Computerization of the Archivo General de Indias: Strategies and Results*, (Washington, D.C: CLIR publication 76, 1999).

34 The Music Performance Research Centre has digitised 1800 musical performances captured during the last sixty-five years; a collection which in digital form truly enables performance studies (e.g. comparison of live and studio performances). Access to and use of MPRC recordings are restricted by the single issue, which will pose the major obstacle to the future of scholarship: intellectual property rights (IPR). Heritage Lottery Fund Project (DG-95-00980).

35 Anne R Kenney, (1996), 'Conversion of Traditional Source Materials into Digital Form' in David Bearman (ed.) *Research Agenda for Networked Cultural Heritage* (Santa Monica: Getty Art History Information Program, 1996), 41-47. Anne R Kenney, 'Digital to Microfilm Conversion: A Demonstration Project 1994-1996 (Final Report to the National Endowment for the Humanities PS-20781-94, 1998). (http://www.library.cornell.edu/preservation/com/comfin.html).

36 J Moffett, 'The Beazley Archive - Making A Humanities Database Accessible To The World', *Bulletin of The John Rylands University Library of Manchester*, 74.3 (1992), 39-52. (http://lannes.ashmol.ox.ac.uk/Pottery/Script/ DatabaseIntroduction.htm) See also the Lexicon of Greek Personal Names (LGPN) (JC. Moffett, A Case of Data Metamorphosis: a description of the Computerization of LGPN Data, 1996

(http://www.lgpn.ox.ac.uk/itindex.html), which is providing scholars with the tools to reshape our knowledge about the Greek world.

37 The SCAN project is digitising some 400,000 wills and testaments from the Middle Ages until the late 19th century, a resource of some 3 million pages; a valuable starting point for research both by scholars and the general public, especially for genealogists who are among the largest users of archives. George MacKenzie, 'Digitising the Scottish Wills', Terry Coppock (ed), *Making Information Available in Digital Format: Perspectives from Practitioners*, (Edinburgh: The Stationery Office, 1999), 139-149.

38 S. W. Thomas, K. Alexander, and K. Guthrie, 'Technology Choices for JSTOR Online Archive', *Computer*, February 1999, 60-65; see also http://www.jstor.org/.

39 http://moa.umdl.umich.edu/

40 Hans-Heinrich Ebeling and Manfred Thaller (eds), *Digitale Archive: Die Erschließung und Digitalisierung des Stadtarchives Duderstadt*, (Göttingen: Max-Planck-Institut für Geschichte, 1999). This project has made available the archives of a small German town.

41 G. Cathaly, 'Mass Digitisation Production Chain At The Bibliothèque Nationale De France', *Digitisation of Library Materials: Report of the Concentration Meeting & Workshop*, Luxembourg, 14.12.98 (1998), 15- 19. (The article was at (http://www.echo.lu/digicult/en/digit.pdf) in late November 1999, but on 30 April 2000 it was no longer there and the new pointer did not work).

42 D. Knowles, *Great Historical Enterprises*, (London: Thomas Nelson & Sons Ltd, 1963). Of course many of these products were themselves printed on acidic paper and are now the subject of retroconversion efforts. A benchmark study of the impact of computers on the humanities and social sciences is: T. Coppock (eds), *Information Technology and Scholarship: Applications in the Humanities and Social Sciences*, (Oxford: The British Academy, 1999).

43 If WIPO succeeds in establishing that facts in databases should benefit from IPR protection there will be significant disadvantages for teaching and research.

44 (http://www.ornl.gov/TechResources/Human_Genome/home.html). These databases themselves become the raw material for future researchers. For instance, Simon Kasif in 'Datascope: Mining Biological Sequences' (*IEEE Intelligent Systems*, 14.6(1999), 38-43.) describes an approach for assisting in the discovery of 'the fundamental connections between genetic sequences and functions of living organisms.'

45 W.C. Barker, J.S. Garavelli, D.H. Haft, L.T. Hunt, C.R. Marzec, B.C. Orcutt, G.Y. Srinivasarao, L.S.L.Yeh, R.S. Ledley, H.W. Mewes, F. Pfeiffer, & A. Tsugita, 'The PIR-international Protein Sequence Database', *Nucleic Acids Research*, 26.1(January 1998), 27-32. See also (http://wwwnbrf.georgetown.edu/).

46 Seamus Ross, 'Designing a Tool for Research in Disciplines Using Multimedia Data: The Romanesque Sculpture Processor,' in Bocchi, F. & Denley, P. (eds.), *Storia & Multimedia*, Proceedings of the Seventh International Congress of the Association for History and Computing, Bologna 29.8-2.9.1992, (Bologna, 1994), 629-635. CRSBI is creating a digital record with associated text-base of the 100,000 surviving examples of Romanesque sculpture in Britain and Ireland.

47 Images of England is creating a digital image for each of the 360,000 listed structures in the England and linking them to the text-based database listed building record (http://www.imagesofengland.org.uk). The project, which will have cost more than four million pounds by the time it is finished in four years time, will provide public access to a national visual record of the listed buildings of England. This visual record will provide a benchmark for future study of our built heritage and help us to chart changes in it. It will itself be an important cultural resource.

48 (http://intarch.ac.uk Heyworth), M., Ross, S., & Richards, J, (1996), 'Internet Archaeology: an electronic journal for archaeology', in H. Kammermanns (eds), Interfacing the Past. Computer Applications and Quantitative Methods in Archaeology, CAA95, in *Analecta Praehistorica Leidensia*, no.28. (Leiden), 517-523 (http://intarch.ac.uk/news/caa95.htm). Seamus Ross, 'Preservation and Networking in Aid of Research', T. Coppock (ed), *Information Technology and Scholarship: Applications in the Humanities and Social Sciences*, (Oxford: Oxford University Press, 1999), 316-317.

49 This may be for such simple reasons as lack of storage space or legal reasons including those relating to software licensing.

50 R. E. Barry, 'Electronic Document and Records Management Systems: Towards a Methodology for Requirements Definitions', *Information Management and Technology* 27.6(1994), 251-256. See also (http://www.barry.com) for further articles on this issue and links to other resources.

51 Ross, 1998, 7-17.

52 Estimating the size of the web environment is difficult. In its July 1999 Internet Domain Survey the Internet Software Consortium identified some 56.2 million

Internet hosts. In 1999 Network Solutions registered roughly 5.3 million domain names.

53 see J.R. Baron, 'E-mail metadata in a Post-Armstrong World', 1999, (http://computer.org/conferen/proceed/meta/1999/papers/83/jbaron.htm)

54 In 1956 disk storage cost $200,000 a megabyte, by 1991 that cost had fallen to $5 a megabyte, and by 1998 it had plummeted to 5 cents a megabyte. The costs continue to fall, but storage brings with it numerous management costs.

55 Numerous contemporary mining examples using longitudinal data indicate the kinds of results that can be obtained. For instance W.J. Mackillop, S. Zhou, P. Groome, P. Dixon, B.J. Cummings, C. Hayter, and L. Paszat, used radiotherapy data collected over an eleven year period as part of the recordkeeping functions of hospitals to examine how its use had changed ('Changes in the use of radiotherapy in Ontario 1984-1995', *International Journal of Radiation Oncology Biology Physics*, 44.2 (1999), 355-362).

56 S.Ross, 1993. `Historians, Machine-Readable Information, and the Past's Future', in Ross, S. and Higgs, E. (eds.) 1993. *Electronic Information Resources and Historians: European Perspectives*, (St. Katharinen: Halbgraue Reihe zur historischen Fachinformatik, A20 [ISBN: 3-928134-95-7]), 1-20; S. Ross, 'Consensus, communication, and collaboration: fostering multidisciplinary cooperation in electronic records', in *INSAR (Supplement II), Proceedings of the DLM-Forum on electronic records*, 1997, 330-336.

57 The 1999 DLM Forum set in motion a programme to raise awareness among the industrial and software development communities of these issues and to build collaborative initiatives. (http://www.dlmforum.eu.org/).

58 For example, D.R. Adams, D.M. Hansen, K.G. Walker, and J.D. Gash, in 'Scientific Data Archive at the Environmental Molecular Sciences Laboratory', (*Sixth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies* and Fifteenth IEEE Symposium on Mass Storage Systems, March 23-26, 1998) reported on the systems they were developing to address the terabytes of data storage in multiple formats that scientists were generating in research laboratories they supported.

59 Michael Macedonia, 'Why Digital Entertainment Drives the Need for Speed', *Computer*, 33.2 (2000), 124-127. The relatively new gaming and edutainment sector employs 29% of the staff in motion picture business, yet it generates 79% of the revenues (*ibid.,* 126).

60 Held in 1993 and sponsored by the British Academy, the British Library and the Association for History and Computing.

61 D. Pyle, *Data Preparation for Data Mining*, (San Mateo, CA.: Morgan Kaufmann, 1999); J. Han, 'Towards on-line Analytical Mining in Large Datasets', *ACM SIGMOD Record* (March 1998), 97-107.

62 P. Maes, 'Agents That Reduce Work and Information Overload', *Communications of the ACM*, 37.7(1994), 30-40; D. Mladenic, 'Text-Learning and Related Intelligent Agents: A Survey', *IEEE Intelligent Systems*, 14.4 (1999), 44-54.

63 As Maxwell M. Burnet and Robert M. Supnik note in their 1996 article 'Preserving Computing's Past: Restoration and Simulation,' (*Digital Technical Journal*, 8.3; (http://www.digital.com/info/DTJN02/), 'The implementation of a particular simulator begins with collecting reference manuals, maintenance manuals, design documents, folklore, and prior simulator implementations for the target system. This is nontrivial. In the early days of computing, companies did not systematically collect and archive design documentation. In addition, collected material is subject to information decay, as noted earlier. Lastly, the material is likely to be contradictory, embodying differing revisions or versions of the architecture, as well as errors that have crept in during the documentation process.'

64 Ross and Gow, 1999, iv-v.

65 Doron Swade, `Collecting Software: Preserving Information in an Object-Centred Culture', in Seamus Ross and Edward Higgs (eds.), *Electronic Information Resources and Historians: European Perspectives*, (St Katharinen, 1993), 93-104.

66 Michael Wettengel, 'German Unification and Electronic Records', Edward Higgs (ed), *History and Electronic Artefacts*, (Oxford: Clarendon Press, 1998), 264-276. It is worthy of note that the CIA had obtained the only copy of one of the key files needed to reconstruct these records, but this file was useless to them without access to other files held only by the Germans. The files held by the Germans were on the other hand of little value without the key file held by the CIA.

67 Seamus Ross, 'Responding to the Challenges and Opportunities of ICT: The New Records Manager', Adrian Allen (eds), *Professionalism Plus*, (Business Archives Council, Proceedings of the Annual Conference, 1998), 9-25.

68 In many ways even more worrying is the debate as to whether records should remain in the custody of their creating organisation or be transferred to archives. See the National Archives of Australia.

69 Ross 1998. See also T. Eastwood, Shadrack Katuu, JKillawee, and JWhyte, 'Appraisal of Electronic Records,' Maria Guercio (ed) *Archivi per la Storia, (Rivista dell'Associazione Nazionale Archivistica Italiana*, 1999), 277-300.

70 Not all digital resources will have their associated documentation in digital form (e.g. manuals). While it remains on paper it can easily be severed from the digital resources which it describes and is susceptible, even under optimal conditions, to loss. In order to mitigate against the loss of information that might occur if these documents became unavailable, strategies and procedures will need to be established to ensure that they are converted into digital form. The research into the use of XML (see T. Usdin and T. Graham, 'XML: Not a Silver Bullet, But a Great Pipe Wrench,' *StandardView,* 6.3 (1998), 125-132) to aid the preservation of digital resources is increasingly extensive: see for instance (http://www.icpsr.umich.edu/DDI/intro.html) or for the central place XML encoding of metadata in work of the Object Oriented Data Technology (OODT) group which carries out research into object oriented data systems technologies for Section 389, JPL, and NASA. Their report *Object Oriented Data Technology for Interferometry Systems* (http://oodt.jpl.nasa.gov/doc/ reports/annual/1999/) provides an example of the key role that XML can play in preservation and access strategies. Currently the conversion to eXtensible Markup Language (XML) offers the best preservation option. Where it proves impossible to convert this material into XML format the preferred digital format for these materials should be Tagged Interchange File Format (TIFF), which is recognised as a preservation standard.

71 See note 2 above.

72 The alphanumeric string 'SW1 8NR' would be meaningless if you did not know that it was a postcode and postcodes were used to refer to particular geographic areas in Britain. The data 'SW1 8NR' is meaningless without other data that provides context, definition, and helps you to process the primary data.

73 National Library of Australia, Preservation Metadata for Digital Collections, 15 October 1999, (http://www.nla.gov.au/preserve/pmeta.html) or National Archives of Australia, *Recordkeeping Metadata Standard for Commonwealth Agencies*. Version 1.0. May 1999 (http://www.naa.gov.au/govserv/techpub/ rkms/intro.htm). Barbara Reed, 'Metadata: Core record or core business? *Archives and Manuscripts*, 25.2 (1997), 218-41; See also the summary of the Third Metadata Workshop (http://www.echo.lu/libraries/en/metadata/ metadata3.html). Michael Day and Andy Stone, 1999. 'Metadata: The Third Luxembourg Metadata

Workshop, *Ariadne,* 20, (http://www.ariadne.ac.uk/issue20/metadata/).

74 For instance the work in digital preservation metadata, *Metadata Specifications Derived from the Functional Requirements: A Reference Model for Business Acceptable Communications*, 9/18/96, (Functional Requirements for Evidence in Recordkeeping: The Pittsburgh Project), (http://www.lis.pitt.edu/~nhprc/meta96.html) was never comprehensively tested. However the issues it tackled and how it tackled them have shaped much subsequent work in preservation metadata in the archives and libraries areas.

75 Seamus Ross, 'Commentary on the Pittsburgh University The Recordkeeping Functional Requirements Project: A Progress Report', *The Society of American Archivists 1995 Conference*, Washington D.C. August 31 1995, (http://www.hatii.arts.gla.ac.uk/sross/saa1995.html).

76 Ian Macfarlane, 'Electronic Records Management Systems, A Report from the Working Group', *Electronic Access: Archives in the New Millennium*, Proceedings, 3-4 June 1998, (London: Public Record Office, 1998), 70-73.

77 There are a number of excellent bibliographic sources, including: (1) Preserving Access to Digital Information, (http://www.nla.gov.au/padi/) (2) United Kingdom Office for Library Networking, (http://homes.ukoln.ac.uk/~lismd/preservation.html) (3) the University of Pittsburgh Functional Requirements Project, (http://www.lis.pitt.edu/~nhprc/bibtc.html) (4) the InterPARES project-American Team, (http://is.gseis.ucla.edu/us-interpares/bibgraph.htm).

78 See for instance Titia van der Werf-Davelaar, 'NEDLIB: Networked European Deposit Library', (http://www.exploit-lib.org/issue4/nedlib) or the NEDLIB site itself, (http://www.konbib.nl/nedlib/).

79 see for example the Data Documentation Initiative.

80 The InterPARES Project, (http://www.interpares.org); Kristine L. Kelly, Alan Kowlowitz, Theresa A. Pardo, and Darryl E. Green**,** *Models for Action: Practical Approaches to Electronic Records Management & Preservation*, (Albany, CTG Final Project Report CTG 98-1) July 1998. (http://www. ctg.albany.edu/resources/pdfrpwp/mfa.pdf). There are numerous other initiatives such as the European Commission funded work on functional requirements for metadata (see Ken Hannigan, 'National Archives and Electronic Records in the European Union', Luigi Sarno (ed), *Authentic Records in the Electronic Age: Proceedings of an International Symposium*, Vancouver: InterPARES Project & Istituto Italiano di Cultura Vancouver, 2000, 36-52, esp 46-

47) and work at pharmaceutical companies such as Pfizer and Astra.

81 Charles M Dollar, *Authentic Electronic Records: Strategies for Long-Term Access*, (Chicago, 1999).

82 A search of the web produces many computer preservation groups, museums and research projects. See also Roger Bridgman, 'What's in a Computer?', Suzanne Keene and Doron Swade (eds), *Collecting and Conserving Computers*, (London: The National Museum of Science and Industry, 1994), 7-18. In its 1986 report the Committee on Preservation of Historical Records noted: 'Moreover it must be realised that no archival organisation can hope realistically to maintain such hardware itself. Integrated circuits, thin film heads, and laser diodes cannot be repaired today, nor can they be readily fabricated, except in multimillion-dollar factories' (page 68). There is one area where preservation of hardware is essential and this is the case of peripheral devices. While it is relatively simple to write a device driver to connect a peripheral device to computer it is much more difficult to build a tape or disk reader from scratch.

83 Margaret Hedstrom, 'Research Issues in Migration and Long-Term Preservation,' *Archives and Museum Informatics* 11 (1997), 287-291.

84 Ross and Gow, 1999, 27-36. (http://www.hatii.arts.gla.ac.uk/Projects/BrLibrary/index.html).

85 See British Standards Institution (99/621800 DC), *Bundles for the perpetual preservation of electronic documents and associated objects*.

86 R Moore, C. Baru, A. Gupta, B. Ludaescher, R. Marciano, and A. Rajasekar. *Collection Based Long Term Preservation.* San Diego Computer Computer Center, San Diego, June 1999. (http://www.sdsc.edu/NARA/Publications/nara.pdf/).

87 Ross and Gow, 1999, 37-38; C. Cifuentes and M. Van Emmerik, 'UQBT: Adaptable binary translation at low cost', *Computer*, 33.3 (2000), 60-; C. Cifuentes and V. Malhotra, 'Binary translation: static, dynamic, retargetable? *Proceedings of the 1996 IEEE Conference on Software Maintenance*, (1996), 340-349.

88 This is not primarily because of media unreliability, because, as Ross and Gow (1999) have noted, magnetic media is durable even under extreme conditions. Although media decays and because of particle breakdown loses its signal, the main long term difficulty is likely to be peripheral device obsolescence.

89 For instance Ken Thibodeau has argued, in a paper presented at the European Commission sponsored DLM-Forum in October 1999, ('Persistent Object Preservation: Advanced Computing Infrastructure for Digital Preservation')

that NARA is now in a position to accession successfully substantial record collections (e.g. one million diplomatic records annually and 25 million messages from the Clinton administration). Reagan Moore, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta, 'Collection-Based Persistent Digital Archives - Part 1', *D-Lib Magazine* 6.3 (March 2000) ISSN 1082-9873 (http://www.dlib.org/dlib/march00/moore/03moore-pt1.html). Reagan Moore, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta, 'Collection-Based Persistent Digital Archives - Part 2', *D-Lib Magazine* 6.4 (April 2000) ISSN 1082-9873 (http://www.dlib.org/dlib/april00/moore/04moore-pt2.html).

90 Deborah Woodyard, 'Farewell my Floppy: a strategy for migration of digital information' (1998), at the National Library of Australia website: (http://www.nla.gov.au/nla/staffpaper/valadw.html). The handcrafted nature of migration is abundantly clear from Woodyard's work.

91 Two very simple examples might help: (1) I have files that began life on a Morrow Computer running CP/M in Wordstar (1982), that were moved to an IBM compatible computer running Wordstar for DOS (1985), eventually moved to WordPerfect for DOS (1989), then to WordPerfect for Windows (1993), and then through a variety of Microsoft Word for Windows formats (beginning in 1997), but which remain accessible today (2000). (2) Of course dynamic documents created in Word 6 that included references to Word macros failed to migrate successfully to Word97. The migrated files no longer display origination (e.g. letterhead) data as this only existed in the macro and was instantiated at runtime and Microsoft changed the macro language from one version of Word to the next.

92 T. Hendley, *Comparison of methods and costs of digital preservation*, (London: The British Library, 1998).

93 On the other hand as data and records are moved to new system environments that may provide richer data structures and more system functionality it may prove possible to do new things and uncover new kinds of information from within the data set itself.

94 When such loss occurs it needs to be documented. Inclusion of metadata describing change history forms a critical element in any preservation strategy.

95 Among the current competing preservation metadata models the Cedars model shows promise. (http://www.leeds.ac.uk/cedars). The use of repositories is a proven

strategy and has been adopted, for instance, by the Library of Congress and NARA.

96 CCSDS 650.0-R-1: *Reference Model for an Open Archival Information System (OAIS)*. Red Book. Issue 1. May 1999. (http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html) In order to take advantage of the preservation strengths of the model it is essential that appropriate documentation and metadata standards be adopted. This information must be linked or wrapped round the resources themselves and be encoded in a functional and secure standard. A strategy of this kind ensures that the platform is in place to permit the application of best preservation strategy, whether that is migration, persistent objects or emulation to each digital resource. The focus of this strategy is on flexibility and responsiveness to changing conditions. David Holdsworth and Derek M Sargent, 'A Blueprint for Representation Information in the OAIS Model', (2000), (http://gps0.leeds.ac.uk/~ecldh/cedars/nasa2000/nasa2000.html) demonstrates how the model might work.

97 Jeff Rothenberg, 'Ensuring the Longevity of Digital Documents', *Scientific American*, 272(1) (January 1995), 24-29; Doran Swade, (1993), 'Collecting Software: Preserving Information in an Object-Centred Culture', in Seamus Ross and Edward Higgs (eds), *Electronic Information Resources and Historians: European Perspectives*, (St Katharinen: Scripta Mercaturae Verlag, 1993), 93-104; Seamus Ross (1998), 'The Expanding World of Electronic Information and the Past's Future', in Higgs, E, (ed.), *Historians and Electronic Artefacts*, (Oxford: Oxford University Press, 1998), 6-28.

98 Lorenzo Valla, *De falso credita et ementita Constantini donatione declamatio*, (Mainz, 1518). Others had suspected and argued this was the case.

99 See for instance, C. M. Kelly, 'Later Roman bureaucracy: going through the files', in A.K. Bowman and G. Wolf (eds) *Literacy and Power in the Ancient World*, (Cambridge, 1994), 161-176.

100 Jocelyn Penny Small in *Wax Tablets of the Mind Cognitive Studies of Memory and Literacy in Classical Antiquity* (Routledge, 1997) built some of her arguments on how the environment, materials (e.g. scrolls) and even the desk conditioned ancient thought.

101 Ross and Gow, 1999, 27-36. There are numerous emulators being developed see for example, (1) (http://ei.cs.vt.edu/~history/emulators.html) (2) (http://www.chac.org/chhistpg.html)

102 Jan Van Der Spiegel, 'ENIAC-on-a-Chip', *PennPrintout*, 12.4 (March 1996),

(http://www.upenn.edu/computing/printout/archive/v12/4/chip.html) and (http://www.ee.upenn.edu/~jan/eniacproj.html).

103 There are many emulators for the PDP-11. These include Software Resources International (SRI), CHARON-11(http://www.charon-11.com/), of Ersatz-11, a software PDP-11 emulator for MS-DOS PCs (http://www .dbit.com/). These emulators support a wide range of device drivers.

104 (http://www.dcs.warwick.ac.uk/~edsac/).

105 Jeff Rothenberg made some very significant contributions to the development of thinking in this area. The papers he produced for the Netherlands are illuminating. See for instance 'Carrying authentic, understandable and usable digital records through time' on the website of Digital Longevity: (http://www.archief.nl/digiduur) under 'Bibliotheek'.

106 B. Bhushan and R. M Phelan, 'Overview of Challenger Space-Shuttle Tape-Data Recovery Study', *IEEE Transactions on Magnetics*, 23.5(1987), 3179-3183. C.H. Kalthoff, R.L. Bradshaw, E.A. Bartkus, and B.I. Finkelstein, 'Magnetic-Tape Recovery And Rerecording Of Data', *Journal of Applied Physics*, 61.8(1987), 4004-6.

107 S.B. Robertson, Digital Rosetta Stone: a conceptual model for maintaining long-term access to digital documents. *Thesis (MSc), Air Force Institute of Technology, Graduate School of Logistics and Acquisition Management, 1996.* (http://www.au.af.mil/au/).

108 Ross and Gow, 1999, 24-25. D. Rugar, H.J. Mamin, P. Guethner, S.E. Lambert, J.E. Stren, I. McFadyen, I. and T. Yogi, 'Magnetic force microscopy: General Principles and Application to Longitudinal Recording Media', *Journal of Applied Physics*, 68.3(1990). J.J. Sáenz, N. García, P. Grutter, E. Meyer, H. Heinzelmann, R. Wiesendanger, L. Rosenthaler, H.R. Hidber, and H.-J. Gütherodt, 'Observations of magnetic forces by the atomic force microscope', *Journal of Applied Physics*, 62.10(1987). Recent articles such as 'Extracting media noise characteristics from MFM images' by P. Arnett, T. Minvielle, and S. Nair (*Journal Of Magnetism And Magnetic Materials* 193.1-3 (March 1999), 479-483) indicate the potential of MFM. M. Boyd and X. Xu, 'MR glide inspection for hard disk defect detection', *Proceedings of SPIE - The International Society for Optical Engineering*, 3619, 53-64.

109 C.R. McClure and J.T. Sprehe, 'Analysis and Development of Model Quality Guidelines for Electronic Records Management on State and Federal Websites'

Final Report. (January 1998).
(http://istweb.syr.edu/~mcclure/nhprc/nhprc_title.html).

110 Indeed the NASA adopted the view that 'because all e-mail can be the target of a number of public and legal disclosure instruments, and as the government's definition of 'records' is difficult to interpret and this policy is difficult to enforce, the agency has stipulated that all email files (central store only) that are older than 60 days must be erased automatically.' (Heather Harreld, 'NASA orders all email destroyed', *Federal Computer Week*, 6/2/97,
(http://www.fcw.com/fcw/articles/1997/FCW_060297_487.asp).

111 S. Shulman, *Owing the Future* (Boston: Houghton Mifflin Co, 1999).

112 We should not be tricked into believing that the survival of vast quantities of data will alone provide the fertile soil for future scholarship; research 50 years from now will be very different. The quality of the data (whether texts, multimedia, databases, or audio), the training of scholars, and the tools to investigate the data will each continue to influence the products of research.

113 In 1999 the Joint Information Systems Committee (UK) released a CD-ROM titled, *Advisory Group on Computer Graphics: Reports and Resources Archive.* Shortly after receiving the CD a note followed saying that 'a virus has accidentally been included'. Although the virus does not have a 'destructive payload' it is easy to imagine the dangers that will lurk in the records that we pass to the future.

National Preservation Office
The British Library
96 Euston Road
London NW1 2DB
Tel: 020 7412 7612
Fax: 020 7412 7796
Email: npo@bl.uk
www.bl.uk/npo/

National Preservation Office
supported by
The British Library
The Public Record Office
The National Library of Scotland
Trinity College Library Dublin
The Consortium of University Research Libraries
Cambridge University Library
The National Library of Wales
The Bodleian Library, Oxford University