Using the papers of contemporary British politicians as a testbed for the preservation of digital personal archives

Susan Thomas, *The University of Oxford* and Janette Martin, *The University of Manchester*

Abstract. *Paradigm (Personal ARchives Accessible in DIGital Media) is an exemplar project to explore how archivists might select, acquire, process, store, preserve and provide access to the digital archives of individuals for the use of future researchers. Using the papers of contemporary British politicians as a testbed, the project team will evaluate existing and emerging theoretical and practical work in the fields of archival science and digital curation. We intend to learn from both disciplines and apply this knowledge to our exemplar scenario with the goal of striking a balance between theoretical principles and practical procedures. This article places the Paradigm project in the broader framework of digital preservation initiatives in the UK and abroad, introduces the key aims of Paradigm and outlines some of our initial findings. We also confront the implications of exponential growth in the creation of personal digital collections - from digital images, music files to personal websites and blogs - and conclude with a discussion of what this means for the wider archival profession.*[1]

Introduction

The Bodleian Library in Oxford and the John Rylands University Library in Manchester have long collected the personal archives of significant figures from all walks of life. These figures include among others academics, composers, diplomats, journalists, politicians, scientists and writers. Both institutions are convinced of the value of acquiring and preserving personal archives, and are conscious that if they are to continue collecting they must develop the capacity to manage and preserve hybrid paper and digital archival collections. At Oxford University Library Services (OULS), the catalyst for action on digital archives came in 2003 when OULS appointed a new Keeper of Special Collections and a new Head of the Oxford Digital Library (ODL). Prior to his appointment as Keeper, Richard Ovenden had been involved in several digital initiatives, including a digital preservation research project at Edinburgh University Library where he was Director of Collections.[2] The new Head of the ODL, Michael Popham, was previously Project Manager for the Oxford e-Science Centre and Head of the Oxford Text Archive; he was also involved with the landmark digital preservation project, CEDARS.[3] Understandably, both men are eager to develop the capacity to preserve digital archives in Oxford libraries. Staff at the Special Collections department of the John Rylands Library (JRUL) are also interested in developing digital preservation expertise. In fact, archivists from the Library's Modern Literary Archives Programme

have already begun some practical exploration in the area. In 2002, the Rylands' literary archivists joined forces with their peers elsewhere in the UK to embark on some small-scale experimentation with the preservation of writers' emails. They found that undertaking such exploratory work in their 'free time' severely limited what could be achieved: the work needed more staff time devoted to it, as well as dedicated IT expertise. The problem, at both institutions, was a lack of resource to examine the issue of digital preservation properly.

In April 2004 a potential solution appeared on the horizon. The Joint Information Systems Committee (JISC), issued a call for projects under a programme entitled 'Supporting Digital Preservation and Asset Management in Institutions' and a bid to explore the preservation of hybrid paper and digital personal archives lead by the University of Oxford in partnership with The University of Manchester was one of 11 successful applications. Money, previously secured from Oxford's Research Development Fund, supplemented this grant to enable Oxford and Manchester to finance a two-year project with 2.5 dedicated members of staff.

The wider context

The 'Supporting Digital Preservation and Asset Management in Institutions' programme, commonly, and ironically, abbreviated to 4/04,[4] is an acknowledgement of the growth in the importance and quantity of digital assets which support the activities of Higher and Further Education (HE/FE) institutions. Proper measures to protect the investment made in these assets are required to ensure that they remain accessible for as long as they are of value to the academic community and wider society; indefinitely in some cases. The Collections Grid, devised by the Online Computer Learning Centre (OCLC), is a useful visual representation of the kinds of content that society might wish to preserve; it also models the uniqueness of content types and the level of stewardship required to curate them.

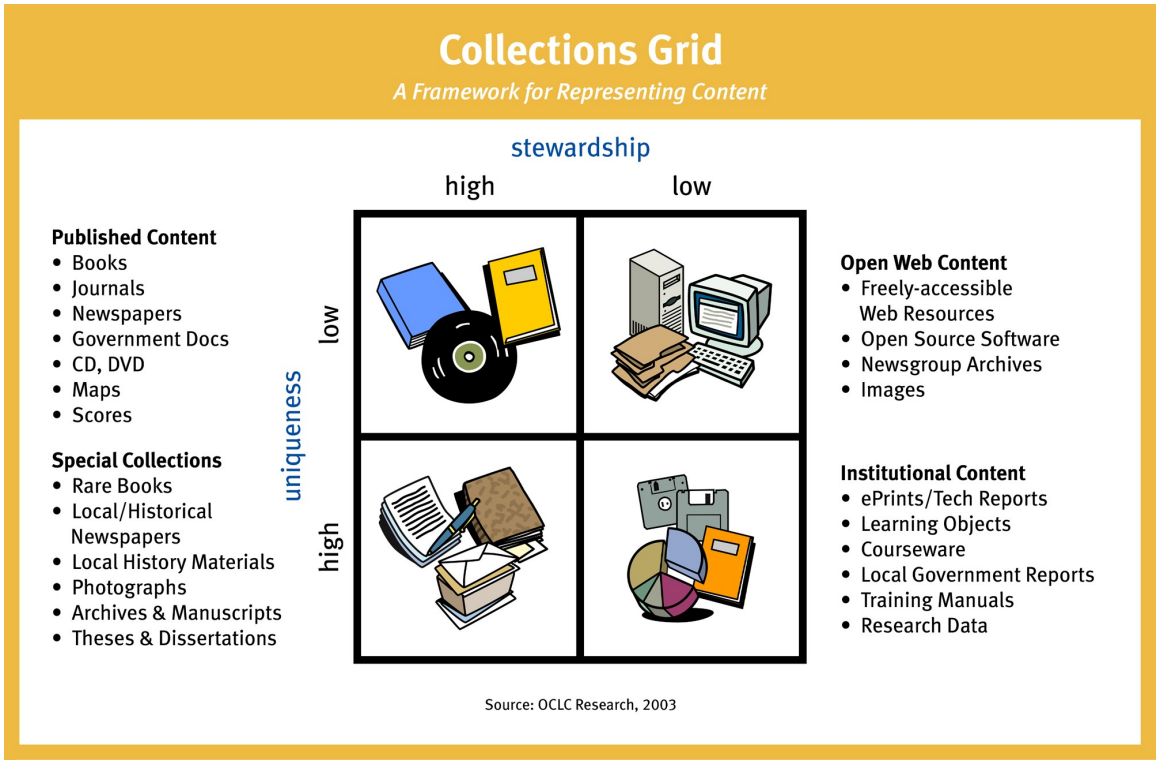

fig. 1: OCLC Collections Grid, 2003[5]

HE/FE communities generate and accumulate most of these different asset types, and more and more of them are produced and consumed in digital form. Given the rapidity and scale of the move from analogue to digital formats, it is not surprising that the 4/04 programme attracted many more project proposals than JISC were able to fund, and that the 11 projects that received funding concern the management and preservation of numerous kinds of asset.[6] These assets include administrative records, such as financial and student records; 'learning objects' such as lecture notes, bibliographies and interactive web-based tutorials; many kinds of research data; and library assets such as e-journals, digitised manuscripts and born digital archives. There has been previous research in all of these areas, but much of its practical application has been geared towards developing workflows, standards and systems capable of creating, acquiring, enhancing, storing and retrieving digital assets; until relatively recently, little attention has been paid to developing preservation functions for such systems. Interest in the preservation aspect of digital curation is growing all the time, especially now that an increasing number of institutions have tools to share and practical experiences to relate.[7]

Preserving digital assets is a concern for all kinds of organisations and the community that has evolved to meet this challenge is heterogeneous and international. Digital preservation is far too complex, and urgent, an issue for any one organisation to tackle alone; therefore, co-operation and standardisation have become watchwords for a digital preservation community, which is characterised by consortia, alliances, networks and partnerships and populated with data creators, curators and (re-)users from all sectors. The establishment of the Digital Preservation Coalition (2001) in the UK and the

Electronic Resource Preservation and Access Network (ERPANET, established in 2001) in Europe were particularly important in bringing people together to discuss their experiences.[8] Further examples of co-operation include collaborations between curating institutions and technology companies, such as the DSpace project of MIT Libraries and Hewlett Packard and the partnership of the Koninklijke Bibliotheek and IBM, as well as consortium projects, such as CEDARS (1998-2005), CAMiLEON (1999-2003), INTERPARES (1999-2006) and the UK Web Archiving Consortium.[9] More recently, the Digital Curation Centre established the Associates Network as a means of connecting individuals.[10]

Reference Model for an Open Archival Information System (OAIS model)

The high-level reference model developed by the Consultative Committee for Space Data Systems: the Reference Model for an Open Archival Information System, usually referred to as the OAIS model or ISO 14721:2003, has been widely accepted by the digital preservation community as a key standard.[11] The OAIS model deliberately eschews jargon from both the IT and archival professions, effectively making both groups speak the same language. Although the product of space data curators, the OAIS model is designed to be as context-neutral as possible. It sets forth a common framework and vocabulary, which is now being used as a planning tool for new digital repositories and as a benchmark for evaluating the capabilities of more established services.[12] The use of OAIS as a benchmark for digital archives may be formalised in the near future as, along with another key document, the Research Library Group's (RLG) *Trusted Digital Repository: Attributes and Responsibilities* (2002), OAIS forms the basis of a model for digital repository certification devised by a RLG and National Archives and Records Administration (NARA) task force.[13] OAIS also serves as a framework for developers of digital repository software; such repositories include DSpace and Fedora, which the Paradigm project is testing.[14] The prevalence of the OAIS model facilitates discussion with those within the digital curation community who have had the opportunity to learn its language, though for the uninitiated, OAIS terminology is a barrier to understanding much of what digital curators are proposing.

This article is not the place for a full-scale introduction to the OAIS model,[15] but, as the OAIS model informs the way digital curators conceive digital repositories, a short explanation of the basic concepts is required. Put simply, an OAIS is:

> an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. [Where] the information being maintained is deemed to need *Long Term Preservation*, even if the OAIS itself is not permanent.[16]

The OAIS has relationships with three entities: *Producers*, which deliver material to the OAIS, *Consumers*, which obtain material from the OAIS, and *Management*, which is responsible for managing the OAIS. The actors in each entity may be human, machine, or both. In addition to defining the entities involved in the long-term preservation of digital materials, OAIS provides an information model for managing the digital materials as they pass through the system. This model consists of three kinds of *Information Package (IP)*, with each IP consisting of the digital object(s) together with the metadata required at that point in the system; these IPs are known as *Submission Information Packages (SIPs)*, *Archival Information Package (AIPs)* and *Dissemination Information Package (DIPs)*. At

the SIP stage, the metadata is supplied by the Producer; this could be the original creator of the material, or perhaps another digital repository. It is likely that the metadata will lack structure and may not be comprehensive at all levels of the archive. At the AIP stage, the SIPs are prepared for preservation; the digital materials submitted for preservation, known as *Content Data Objects*, are combined with the *Preservation Description Information (PDI)* needed to administer the preservation of the object. OAIS breaks the PDI down into four sections: *reference* (a unique identifier), *context* (relationship to other objects), *provenance* (history of the archived object) and *fixity* information (demonstration of authenticity). OAIS also requires the archive to maintain the *Representation Information* required to render the object intelligible to its designated community – this might include information regarding the hardware and software environment needed to view the *content data object* or a look-up table for a database. Recently, a RLG/OCLC working group published the PREMIS data dictionary, which more formally defines the 'things that most working preservation repositories are likely to need to know in order to support digital preservation' in semantic units.[17]

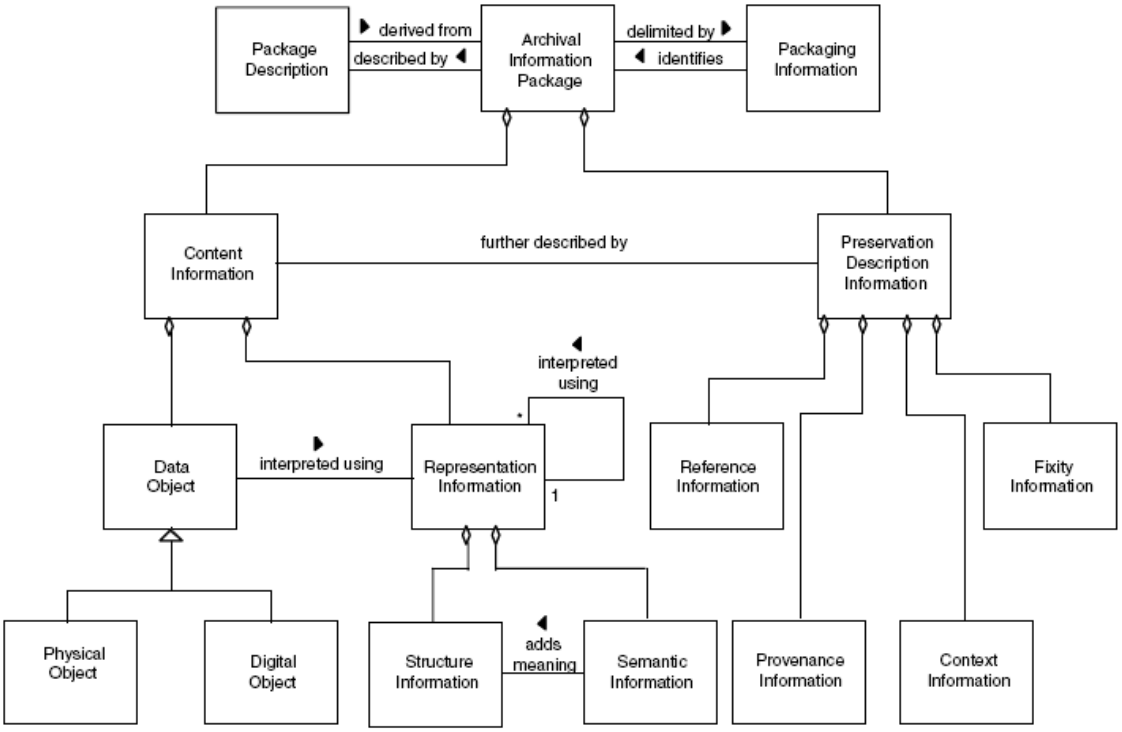

fig. 2: Detailed view of the OAIS model's
Archival Information Package (fig. 4.1, OAIS)[18]

When an OAIS wants to release some of its material to a Consumer, it supplies it as a DIP; the metadata accompanying the object at this stage will be dependent on the Designated Community, but it is likely to be more descriptive than technical. The METS schema has been designed to facilitate this Information Package relationship between objects and their metadata; an XML metadata standard capable of embedding or linking to external XML encoded metadata, such as EAD 2002, METS is being adopted by many digital library projects and is already supported by some digital repository software.[19]

OAIS also provides a functional model, which consists of the following seven functions: *Ingest, Archival Storage, Data Management, Administration, Access, Preservation planning* and *Common Services*; plus information about the kind of activities undertaken by each function. Most of these functions are easily identifiable, but perhaps it is worth mentioning that *Ingest* equates, roughly, with the archival processes undertaken when an archive is newly accessioned and before it is added to archival storage;[20] and that *Common Services* are those required by any IT system, such as the timely application of security patches.

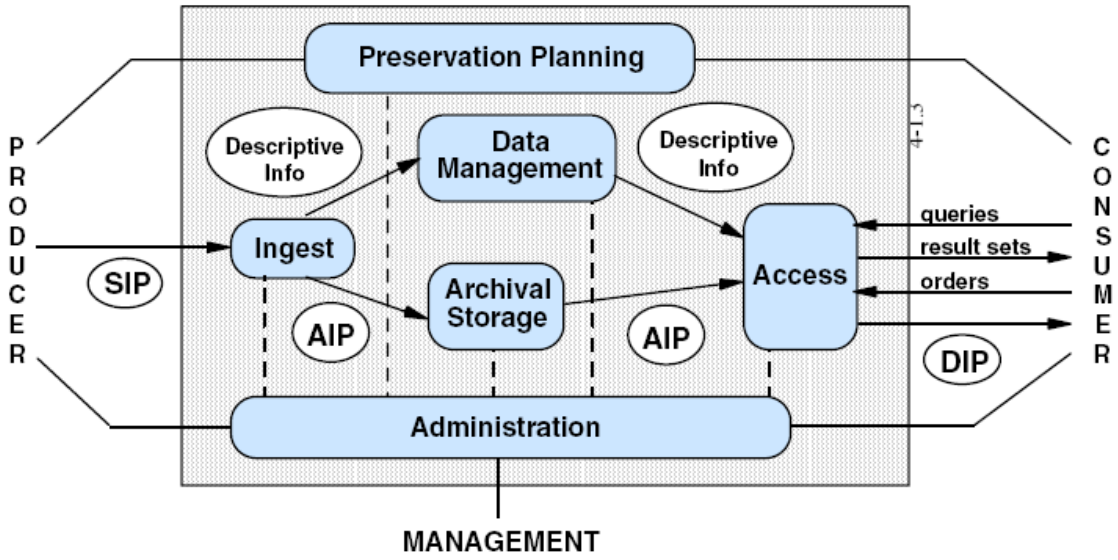

Fig. 3: The OAIS Model's Functional Entities (fig. 4.1, OAIS)[21]

By implementing the models specified in the OAIS standard, it is hoped that organisations will be able to demonstrate compliance with the responsibilities of an OAIS. These responsibilities are defined as follows: negotiate and accept information from Producers; determine which community should become the Designated Community; ensure that Information Packages are independently understandable; ensure IPs are preserved; and make preserved IPs available. Although couched in different language, these responsibilities are recognisable. Archivists already undertake these responsibilities on behalf of creators and users, or 'Designated Communities'. OAIS is about providing an intellectual framework, which will aid us in meeting these responsibilities in a digital environment.

The OAIS concept of 'designated community' means that while digital preservation benefits from the input of various sectors, much of the research and development is focused on developing solutions for specific contexts, giving rise to digital repository types. Amongst these repository types are systems developed by librarians to preserve e-journals. One important initiative in this area is the open source LOCKSS system (Lots of Copies Keeps Stuff Safe) developed by Stanford University Libraries. By working closely with publishers, Stanford's system enables participating libraries to own copies of the e-journals that they subscribe to as well as content published in open-access e-journals.

Each library in the LOCKSS network hosts an inexpensive machine running LOCKSS software, which crawls the websites of e-journal publishers to collect content; by talking to LOCKSS machines at partner libraries, the software engages in a peer-to-peer audit that identifies and repairs any corrupted content using a voting system. LOCKSS only provides libraries with access to content that they have paid for, so it satisfies the needs of publishers. The success of LOCKSS is affirmed by the number of participating libraries, over 80 on five continents, and the number of subscription publishers involved, currently over 60.[22] Another type of repository is the institutional repository, designed to store, manage, and sometimes preserve digital content produced by HE departments.[23] A related trend is the electronic-theses repository; one example of current research in this area is the EThOS project, which aims to develop a prototype UK online e-theses service, which will be easily scalable and financially sustainable.[24] There are also centralised, often discipline-oriented, research data repositories, such as those managed by the Arts & Humanities Data Service (AHDS).[25] Amongst the first to archive websites were the National Library of Australia's PANDORA project and Brewster Kahle's Internet Archive (both established in 1996), but in the last few years, many national libraries have initiated web-archiving activities, and in 2003, many of them formed the International Internet Preservation Consortium.[26] Also important for the future of digital preservation are software repositories; the most popular open source software repository is Sourceforge.[27] Moving toward the archival sector, the vast majority of relevant work is taking place in institutions responsible for preserving national archives. In the UK, the 1999 'Modernising Government' White Paper set targets for government agencies to create and manage their records electronically. Those deemed archival amongst these records will eventually come to The National Archives for preservation in its Digital Archive.[28] Other 'national' institutions investing heavily in preserving born-digital archives include the Koninklijke Bibliotheek, the National Library of New Zealand, the National Library of Australia, the National Archives of Australia, and the NARA.[29]

What most of the initiatives described above have in common is that the curator is dealing with the assets of a 'designated community' that their organisation, or 'community', may lay some claim to. The basis of the claim may differ slightly in each case: for national archives it is underpinned by public record legislation; for serials librarians it is the right of libraries to ensure permanent access to e-journals they subscribe to; for universities, institutional repositories act as a research portfolio and a means to safeguard institutional investment in employee-produced content; for discipline-based research data repositories, the donors and users are often members of the same designated community. In many cases the curator is working with recognisable communities and organisations; further, in some communities the curators have a mandate to influence the workflow and practices of those creating digital assets.[30] Working with the creators of personal archives is entirely different: it entails working with a host of diverse people, cultures, and systems. We collect material which individuals have no obligation to give us; we cannot impose standards governing the creation, management, and disposition of personal archives. We can advise potential donors, but ultimately we cannot compel anyone to follow any recommendation we might make. Unlike organisational records, the management of personal records cannot be driven by legislation or company policy. Collecting institutions, which have traditionally adopted a *laissez-faire* approach to acquisition, compound the situation.[31] All too often archivists in such institutions assume a passive role in which they receive

material at the end of its active life, sometimes long after. Collecting archivists tend to distance themselves from the process of records creation and management, which is often viewed as the remit of the records manager, and in the case of personal archives the remit of the record creator alone. For all these reasons, personal papers have been neglected by digital preservation research to date. This is a significant problem for institutions like the Bodleian and the Rylands. Personal papers are increasingly born digital; many are not intended to have an analogue equivalent or an existence beyond the screen of a computer. Word-processing software threatens the survival of the draft, making it nigh-on impossible for researchers to trace the evolution of a writer's or scientist's thoughts, and, since the arrival of email, instant messaging and texting, letter writing has gone into decline. Paper diaries, address and note books are giving way to personal digital assistants and mobile phones. Whilst the vulnerability of personal digital material is gaining more media exposure, because it is an issue that most of us need to grapple with, cultural organisations cannot rely on the media to reach potential depositors, or to communicate the right messages. This we must do for ourselves. [32] Unless archivists develop the necessary expertise and infrastructure, and work with relevant record creators, historians and biographers will be denied a rich source of material.

The Paradigm project

Paradigm is an exemplar project which is exploring the cultural, legal and technical issues involved in the long term preservation of digital private papers by engaging with record creators and employing sample collections to practice archiving digital private papers. The project, which began in January 2005 and is scheduled to finish at the end of February 2007, is processing materials using traditional archival procedures in tandem with workflows suggested by the OAIS model, with the intention of developing protocols which harmonise the two approaches. It was decided that the project would use the papers of contemporary politicians as its exemplar collections because bringing digital preservation to the attention of politicians is a valuable exercise in itself, and because politicians' archives are well-represented among the personal archives held at both institutions. At the Bodleian Library, researchers can study the papers of six Prime Ministers, over 100 MPs, as well as the Conservative Party Archives. In Manchester, the JRUL holds the papers of Ramsay MacDonald, first Labour Prime Minister, and the papers of several other labour and trade union activists. The JRUL also works closely with the Manchester-based Labour History Archives and Study Centre which cares for the Labour Party Archive.[33] Building on existing strengths, and connections, it was agreed that Oxford would work with Conservative politicians and Manchester with Labour politicians. It was felt that by spreading the project over two archival institutions and at least two political parties, the outcomes of the exemplar would be more representational.

The early part of the project centered on the archivists familiarising themselves with the people, organisations, projects, literature and tools involved in digital asset management; and exploring the less technical aspects of preserving digital private papers. These aspects included re-visiting what we understood by personal archives, selecting politicians to approach; developing relationships with the private offices of politicians; exploring cultural and legal issues; drafting terms of deposit; and making accessions. Subsequently we have made inroads into the more technical aspects of the Paradigm

project, including an exploration of the Fedora open source repository. We have also opted to evaluate the DSpace repository. Both DSpace and Fedora have established and expanding user-bases in the Higher Education and Library sectors.[34] The project staff have also begun to experiment with all manner of tools which might assist the archivist in acquiring, managing, preserving and disseminating digital materials.

A particularly important tool for digital curators is a metadata extractor. Digital objects cannot be left on shelves while we find money to catalogue them. It's possible that the media may survive twenty years sitting on a shelf, we might even be lucky enough to own a device that can read the media twenty years hence, but will we also have the hardware and software necessary to render the file from ones and zeros to something understandable by those of us unschooled in binary? It is crucial for the survival of digital objects that accurate technical metadata is produced in a timely, and economical fashion. We need to know what we have and we need to know sooner rather than later. This knowledge informs the 'Preservation planning' function of a digital repository which helps us to use our limited resources wisely.[35] Hand-crafted metadata is expensive and error- prone; this is why others engaged in digital preservation have developed tools which can examine a file and, if they recognise that file type, can automatically generate the required technical metadata. [36] Paradigm hopes to adapt existing tools to generate metadata designed for the preservation of personal archives.[37] We intend to develop metadata profiles using the METS and PREMIS standards and implement these within a digital repository.[38] We will also be selecting a metadata standard for intellectual property rights (IPRs). Managing IPR metadata is a much higher priority in the digital world because any preservation or access strategy involves copying or redistributing an item and the IPRs of others must be managed over a longer period when collections are accessioned soon after creation.

The project will share what it has learnt about metadata, and other aspects of digital preservation, by developing an on-line best-practice workbook available at the project website http://www.paradigm.ac.uk/. The workbook is intended to be used by IT and archival staff involved in the preservation of digital materials, though we think that Paradigm will be most relevant to collecting institutions, especially those caring for the personal papers of individuals whether they be writers, scientists, politicians or academics. The workbook will include basic guidelines for individuals creating digital records likely to have long-term historical value. Such guidelines will include advice on backup procedures, caring for hardware, 'future-proof' file formats, naming conventions, encryption, using online services, and many other topics. It will also highlight issues relating to various processes involved in digital preservation, provide a glossary to the sea of acronyms, and include template policy and procedural documents. Oxford and Manchester Universities are both committed to maintaining this resource online for 3 years beyond the life of the project, whereupon JISC will assume responsibility for preserving the website.

In addition to learning new technological skills, perhaps one of the exciting aspects of working with politicians and contemporary records is the opportunity to be involved much earlier in the records cycle.[39] Early intervention is an important principle for digital archivists, but it is relatively new to see this kind of relationship between archivists and

creators of personal archives. It turns the archivist's relationship with a depositor on its head. Rather than being approached by a depositor at a time when they are engaged with their memories and their place in history, we are approaching working politicians who may not have considered the historical import of their papers, and are often too busy to pay much attention to the idea. We are making assumptions about the future significance of individuals in the infancy of their careers, or mid-career, when their personal historical significance is not necessarily obvious, though the events and activities in which they are involved may be more so. Simply by selecting an individual to work with, we are conferring significance on them, and by choosing to remember them we are forgetting others. We are undermining what Jenkinson called the 'natural process' underpinning the accumulation of archives.[40] Despite these philosophical issues, the project team decided that the vulnerability of digital records, to accidental or deliberate loss, merited a compromise of principles, and that rather than approach politicians at the end of their careers, we had to be working with them from the beginning to ensure that their personal digital archives survived, in accessible form, for us to curate.

On the advice of its academic advisory board,[41] the project attempted to persuade a range of politicians, at different stages of their careers, to participate in the project. Not all the politicians approached agreed to take part in Paradigm, but the project has certainly succeeded in attracting the variety it sought. To-date, we have worked with members of the Conservative, Labour and Liberal Democrat parties, with peers, MPs and MEPs, and with politicians with international, national and local profiles. Selection for the project has been dependent on a combination of factors: potential historical interest, the willingness of the politician to participate, and the need for our exemplar to address a mix of individuals and contexts. Because the project is primarily of a research and developmental nature, we are acquiring material on fixed-term deposit; this reassured some of our participants who had understandable qualms over the sensitive nature of some of their records. We hope to acquire at least a section of the material at the end of the two-year test bed project for permanent preservation, but this will be subject to renegotiation and another deposit agreement. Realistically, we may not be ready to offer this commitment, or we may have to temper it:

> Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate.[42]

'What *is* personal?'

Working with politicians and their offices has required us to clarify what we mean by personal archives and why we think they are important, if only to explain these things to record creators. Pinning down exactly what is meant by personal archives (private papers, personal records, or manuscripts) is challenging. Naturally, we began by examining existing holdings, identifying the record types found in these, looking for the digital mediums which are being used instead of traditional ones, and thinking about emerging technologies which might have implications for personal digital materials. We were also interested in finding out what roles, activities and relationships our collections bear witness to, as these are also important selection criteria. Examples from our collections show that some personal archives document their creators more comprehensively than others.[43] One of the Bodleian's collections, the personal papers of John Morley, 1st Viscount Morley of Blackburn, comprises a range of personal and professional records: engagement diaries and journals, correspondence with his sister, Grace, and other family

papers; a general correspondence series, and papers originating in his roles as Chief Secretary of State for Ireland, Secretary of State for India; as well as literary papers concerning his *Life of Gladstone*.[44] The Morley papers contrast sharply with those of Eric Heffer, held at the Labour History and Archives Study Centre, which mainly document his political roles and, as such, include record series on: The Communist Party; The Labour Party; The Industrial Relations Bill; Trotskyism and the National Executive Committee chairmanship.[45] Broadly speaking, an ideal archive might document the several roles of an individual: personal and family, professional and other external interests. Such personal archives give readers a unique, human perspective into historical events that is often wanting in the official record:

> Those of us who worked within government know what official records are and they're something very important. They are what is put down for history and they are intended to be defensive against historians, parliamentary questions at a fairly modern date, and they are intended to deal with the feelings of officials. They do not relate very much to what actually happened. I don't mean that they are untrue, and officials would never allow them to be untrue but they are the minimal truth.[46]

The project has not acquired material that documents the personal aspects of politicians' lives. This type of material is of great interest to some historians, but it is difficult to persuade politicians to place this current personal material in a library, especially during a testbed project. Another pertinent issue, particularly during an election year, is a politician's lack of time for preserving private papers. The material obtained so far reflects the professional role of our politicians; the records accessioned include policy briefings, emails, drafts of speeches and other internal documents generated and accumulated by their private offices. The personal archives of politicians are distinctive in that they are not generally the work of one person. Rather they are a joint enterprise with much of the day-to-day correspondence and office papers being created by the MP's personal assistant, or other constituency office staff. Many politicians employ speechwriters which again distances the archival record from the authentic voice of the principal political figure. A sizeable proportion of the 'personal' papers collected to date are circulars from the political party's central office (briefings, research papers etc). For a politician of the governing party, matters are further complicated by the potential for overlap between the content preserved in a personal archive and that contained in official records preserved under legislative requirements by The National Archives. In such instances, the curator of the personal archive would need to refer to The National Archives regarding the classification of similar material. Interestingly, the problem of how we define the 'personal papers' of a politician has also engaged the attention of researchers in Australia who have found that they were also, to a certain extent, collecting the office papers of politicians:

> Once a person acts in an official capacity in an organisation it becomes an issue of drawing the boundaries between the personal records and the records of an organisation. The records from the Minister's office can be conceived of as the records of the Minister, Ms X, the records of the Office of the Minister for Y, or the records of Portfolio Z. The 'official person' is rarely the sole direct creator of the records under his or her immediate control.[47]

The involvement of third parties in personal archives raises a number of issues, and whilst these are not exclusive to the digital domain, the currency of the records magnifies their importance.[48] Some of the politicians participating in the Paradigm project have been circumspect about providing copies of confidential records which could compromise others, such as email, constituent casework records or engagement diaries. Other anxieties include information falling into the wrong hands, either in transit or at the repository.[49] Leaks and negative media coverage are a particular concern, and gauging when it is safe to open records to researchers will be as difficult for those forging their careers, as for those at the height of their power. Participants have therefore required reassurance of our personal and institutional discretion, our policy of keeping private material closed to researchers and our technical competency in ensuring the security of their papers.

One of the earliest tasks for the Paradigm archivists was drafting an appropriate deposit agreement. This was challenging, as it proved impossible to find other examples of deposit agreements drafted for digital personal papers. The media hype surrounding the Freedom of Information Act (2000), which came into force in England, Wales and Northern Ireland in January 2005, heightened fears about the disclosure of private information held by public institutions and these had to be addressed in the deposit documentation. The Act does provide exemptions that can be integrated into deposit agreements, such as s. 41, which provides an exemption for material provided in confidence, but uncertainty is likely to continue until the status of deposited and donated private collections under FOI is clearer. Identifying and protecting Intellectual Property Rights (IPR) and privacy issues were also key considerations when drafting our deposit agreement for several reasons. The first being that digital preservation depends upon the ability to make multiple copies for preservation purposes: this fundamental requirement is stated explicitly in the Paradigm deposit agreement.[50] Other IPR issues included primary and third party copyright within deposited collections. A politician is normally the primary copyright holder in their archive, but their papers may include hundreds of images, some of which will have been created in-house, others could have been forwarded from any number of other creators. Email correspondence raises similar issues. Institutions may have to consider assessing the risk of violating rights where tracing rightsholders is too sizable a task to contemplate. The records of politicians also contain material which falls within the scope of the Data Protection Act. Putting the legal implications of copyright and privacy laws aside, some of our politicians have raised ethical concerns that it is wrong to supply records generated by others, arguing that the creator would not have envisaged this ending for their missive and may not agree with it.

Many of the paper records generated by our politicians are also at risk. Most of our participants are short of office space and it is common practice to destroy old material during the parliamentary recess. After a general election campaign, a change of brief, or the redrawing of a constituency boundary, destruction can be even more extensive. However, once paper records reach the archive their preservation is largely a passive exercise; management decisions regarding appropriate physical storage and access conditions can be applied in blanket fashion. Preserving digital archives is more difficult for many reasons: records are easily duplicated and altered; a record may contain multiple file formats; and there are so many different types of file, each requiring its own preservation strategy and each dependent upon a specific combination of hardware and

software. There is a danger that in the case of personal archives, where no organisational body is present to impose standards or policies relating to digital recordkeeping, let alone implement Electronic Document Management and Records Management (EDRM) systems, that unless archivists accession records soon after they are created, or offer support to record creators to maintain their own digital archiving systems, then they will not survive.[51]

Are we being unnecessarily alarmist? At the Bodleian and the JRUL we are still finding that recent accessions of personal papers are largely paper. This is partly explained by the timing of accessions, which usually occur toward the end of the creators' career, if not posthumously, and which result in the accession of records that are often decades old. In these cases it is unsurprising that much of the material is paper and that readiness for digital accessions does not top the archival agenda when there are so many other issues demanding attention. This paper mentality leads to assumptions that individuals are printing important documents, though the shift towards an increasingly digital culture argues to the contrary. Indeed as IT becomes increasingly sophisticated and the populations' digital literacy grows, one result will be more complex records that do not translate well when printed. Whilst it is tempting to think that we do not really need to worry about preserving digital personal papers, this would be a complacent and blinkered approach to acquisition. We may not be receiving great swathes of digital material, but this does not mean that it does not exist.[52] It is more likely to mean that we need to educate our donors to think of their digital materials as part of their archive. We also need to preserve these materials digitally, to maintain as much of their digital qualities as we can economically justify, and to retain the context of their storage and use. Whilst printing digital objects is one means of preserving them, it entails great compromises: many digital objects lose formatting, relationships, intertextuality, as well as other functionality, when printed. All this is a grave loss to researchers.

One of the key issues facing the Paradigm project is how to manage the appraisal of paper and digital records in tandem. Many people routinely print paper versions of digital records for ease of use and we have found that many of our participants have both hardcopy and digital copies of key documents such as election leaflets and reports to constituents. Even if a politician's office creates all their papers electronically, there will still be some documents, such as letters from constituents, invitations and press cuttings, which are received in a paper format. Records relating to these paper records, perhaps the images used in a document, will also be found within digital systems and the archivist needs to find a way of identifying overlaps between digital and hardcopy records and linking related material. Hybrid record keeping systems risk unnecessary duplication: a practical measure is to audit both the digital records and paper records together, establish where the same documents exist in both mediums and decide which should be retained as the archival copy. Where both paper and digital copies exist, it would seem sensible to treat the digital as the 'master' copy unless the paper copy includes autograph annotations. The digital record has search and manipulation benefits which the paper record cannot equal.

Another major consideration when devising procedures for digital records is establishing mechanisms to preserve the integrity and authenticity of the digital object during the movement from creator to the preservation system, and thereafter. The process of

acquiring digital material as experienced by the Paradigm project is a curious blend of records management, IT and traditional archival skills. The process begins with a records survey in which a questionnaire is sent out in advance to all participants.[53] This is followed by a visit from members of the Paradigm team to introduce the project; answer questions; gather answers to the questions set out in the records survey questionnaire; assess functions, staffing structures and responsibilities; and to appraise the records. During the first visit, screen prints or text files of directory structures from all the office computers holding relevant data are created;[54] this was found to be an effective means of identifying exactly which folders were of interest and conveying this information to the participants. Records to be accessioned can then be agreed between the participants and the project team. Once the scope of the accession is understood, the archivist arranges a visit to make the accession. Equipped with USB sticks, laptop and blank CDs to capture the digital records, the archivist follows a transfer protocol which includes the completion of a transfer form, recording any provisos, such as access restrictions, as well as checksum information to ensure that the material accessed at the repository is identical to that accessioned at the politician's office. Even for the IT savvy it can take a while to orientate oneself on an unfamiliar computer often while holding a conversation with office staff. For this reason, it is essential to gather as much information on the software and hardware being used by the depositor at the survey stage.

Our initial accessioning visits raised a number of technical issues: authenticity, technical validity of formats, viruses, security, and duplication to name a few. We quickly concluded that USB sticks and CD's, while adequate for acquiring small amounts of data, were often too slow when accessioning large and complex data.[55] We are currently testing the use of a portable hard drive installed with a tool kit (virus checker, checksum software and directory structure software). As well as putting strategies in place to deal with these issues, archivists dealing with digital or hybrid accessions will also need to become familiar with the export features of popular software packages and services so that they are able to extract the material selected for preservation in the right formats and preserve as much of the directory structure ('original order' in archival language) as possible. The workbook will include how-tos for some of the technologies we come across, but cannot hope to be comprehensive. For personal papers, the most challenging accessions are likely to be email (for example, obtaining email from a Hotmail account, or exporting from Microsoft Outlook) and exporting data, such as appointments and addresses, from personal digital assistants or mobile phones.

Personal Digital Media – what's on your hard drive?

The rest of the article will look beyond political personal papers and the Paradigm project to broader issues concerning digital media and the personal record. The processing power available to individual consumers is evolving continuously and can support increasingly sophisticated software capable of creating infinitely more complex digital objects. The human instinct to collect is assisted by the evolution of storage technologies that enable us to store more and more, while costs decline: 'There is more room to store stuff than there is stuff to store'. Time and skill are now the only restrictions to generating content.[56] In addition to affordability and growing capacity, data storage is now much more portable and flexible than ever. Portable devices, such as USB keys, portable hard-drives, i-pods

and suchlike are common; and online services offer remote storage, accessible from internet cafes worldwide, or anywhere you can hop onto an unsecured wireless network. This raises the question of how to manage digital personal collections so that the collector can actually find what they need when they need it. Computer scientists engaged in the 'Memories for Life' project estimate that by 2019 'the digital archive of even one person…is likely to consist of petabytes of linked images, documents and audio.'[57] The challenge will be creating indexing strategies that can evolve to meet new demands. Gmail has innovative indexing strategies – rather than filing in directories you add your own 'labels' to email. Metadata can also be added to your images using photo-album software. Software developers are making this kind of indexing available to personal consumers, but it will come as no surprise that software houses have not made the adoption of open metadata standards, which would give customers the freedom to switch to rival companies, a priority. Online services also offer this kind of tagging, but again their metadata is not standardised, so if you wanted to transfer your life's collection of photographs to another service provider, the way you might transfer your bank account, you may be able to get the images out, but not necessarily with their metadata attached. In such a climate, users need to be careful that they do not lock their precious data into these services.[58]

Personal digital material is not just stored on your PC and media in your house, but is also to be found on other people's servers – very different to the boxes in the attic and shoeboxes in the wardrobes. There are a plethora of online services available to individuals which offer tools to create, customise, share and search content. The 'blogosphere' now contains 27.5 million blogs, and, in September 2005, Google Blog Search was launched to search them all.[59] Services for email or images are widespread and the 'Ourmedia' service, launched March 2005, caters for absolutely any kind of digital content you care to create, though it encourages submissions in open formats.[60] Interestingly, many of these services are claiming that they will look after your personal digital material 'forever', but they do not divulge how they intend to achieve this or exactly what they mean. Can we trust these kinds of institutions to honour such commitments, or is this the preserve of established cultural heritage institutions?[61] Could these new services become the cultural heritage institutions of the future?

Given the widespread developments, in relation to the creation and management of digital material, taking place in the internet sector, it is natural to consider whether archivists ought to share, or even hand-on, the mantle for long-term preservation to those who are shaping the future, and those who are already providing the means to store personal digital material. Will Google, and equivalent email providers, become defacto archivists of email because they already hold the content on their servers? It is possible that Google might provide access to email archives in future years, perhaps users will be asked if their email might form part of a future social archive when they sign up for an account. It is very difficult to predict the future especially given the rapid pace of technological change, and the social change it provokes. However if we are to take on the challenge of preserving digital personal papers it seems likely that we will have to sacrifice some of our holy tenets. The archival theory dominant in the UK is best-suited to managing paper records generated by organisations. It evolved from centuries of record keeping that had it roots in diplomatics, land law and a succession of Public Record Acts. The key principles, laid out by Hilary Jenkinson in his *Manual of Archive*

Administration (1922), of provenance and original order have remained a guiding force.[62] Yet the significance of the latter tenet may become weakened in an increasingly digital world in which searches can be performed in the blink of an eye and data can be instantly reconfigured to answer specific queries. Jenkinson would surely disapprove of digital archivists actively engaging with the creators of personal papers and the still greater heresy of seeking to influence how these records are created and stored. Yet if the nation's memory is to be preserved for posterity the era of the impartial passive keeper of records has surely passed.

What do we mean by digital preservation?

After looking at the broader issues concerning the preservation of digital private papers, it might be useful to give a simple introduction to what we actually mean by digital preservation, the 'nuts and bolts' of how it may be possible to preserve the digital record over time. There are several competing theories on how best to preserve digital material, all of which have advantages and disadvantages. Most authorities agree that, where possible, it is vital to retain the original bit stream which can be used as the starting point for subsequent preservation strategies. Beyond this there are two main rival camps; those who believe in migration and those who favour emulation.[63] Three basic migration approaches exist. One approach is to continually migrate obsolete, or near-obsolete, digital formats to newer formats so that the digital object is transferred from one software or hardware generation to the next. Another approach involves the transformation of objects into standard file formats specified by the repository; this approach is sometimes called 'normalisation'. The National Archives of Australia, who convert their digital records into XML, have championed normalisation.[64] Yet another migration option is to migrate as access to individual resources is demanded, rather than migrate on ingest or as formats near obsolescence. The downside to migration is that some of the attributes of the digital object may be lost during the conversion process, for example formatting. The migration method is based on the premise that content is more important than look or feel. Emulation, by way of contrast, keeps the digital object in its original data format but recreates some or all of the original processes enabling the object to be recreated on current computers.[65] Advocates of emulation stress the importance of maintaining the exact look and functionality of the record to be preserved, though it's debatable whether digital materials really have an 'exact' look and feel because they are so dependent on the environment used to render them.[66] Both migration and emulation require a large commitment in resources both upfront and over time.[67] Ongoing migration requires intensive cycles of work to convert objects in obsolete forms to ongoing formats, and all migration methodologies require the development of tools capable of undertaking such migrations on batches of files. Emulation also requires highly skilled computer programmers to write emulator code, and sophisticated strategies to deal with IPR issues that may arise when replicating proprietary software. It seems likely that different file formats will be suited to different strategies. Oltmans' work on digital preservation strategies indicates that the greater the variety of digital objects a repository seeks to preserve, the greater the cost will be, regardless of strategy. Decisions in formalising a strategy will include the relative importance of content and preserving the original experience, the variety of objects which the digital preservation service is expected to preserve, and what kind of batch processing is available.[68]

Given the likely costs of preserving digital records over time, interest in file formats and backward compatibility has grown. As the name implies, open source software (OSS) means that technical information required to understand the software is openly available; users of open source are allowed to run the program, study, modify and redistribute without incurring royalties. This allows software to be modified and adapted to user needs.[69] If the source code is available to future digital curators there is a greater chance that the digital object can be preserved. A team of researchers funded by the Ministry of Defence concluded in 2001 that 'OSS has shown that access to software's source code is a major enabler of flexibility, and hence reduces legacy problems considerably'.[70] Open license applications can spread the development costs across like-minded organisations and their use is gaining popularity in higher education.[71] Users of open source software can customise and extend software and feed the resulting code fed back into the main project where it is made available to others. For example, DSpace, a digital repository software to be tested by this project, is open source and users are encouraged to customise and extend the software. Some commercial software developers also support open source products or are willing to give access to the source code underlying some of their software.[72] Others, such as Adobe, provide access to file format specifications, whilst keeping their proprietary software closed. Adobe has also recently launched PDF/A, a constrained form of Adobe PDF version 1.4, which may simplify the long-term preservation of page-oriented documents.[73]

Unfortunately, PDF/A will not solve all our problems. Individuals create a wide variety of data-types. NARA in Washington believe that some 16,000 software formats are being used throughout the federal bureaucracy, and whilst the number may be smaller for individuals, the variety is endless.[74] The personal digital material accessioned from our participants in the first ten months of the project alone includes some twenty file formats, and this is material created in the past five years. Imagine how many file formats we might use in a life-time. To-date, the project has accessioned over 1000 MB of material, and, in the near-future may be accessioning an email archive containing some 37,000 received emails and a smaller, if considerable number of sent email. Amongst the data-types accessioned so far are email, word-processed documents, spreadsheets, digital images (publicity material), PowerPoint presentations as well as personal web pages and blogs. It is important for Paradigm to deal with the preservation requirements of as many of the different file formats that archivists are likely to encounter as possible.

Conclusions

The cost of digital preservation is likely to be prohibitively expensive but we do not yet have the evidence to make realistic estimates of just how much schemes will cost. NARA awarded Lockheed Martin a $308 million contract to build a permanent archives system to preserve and manage electronic records created by the United States federal government.[75] This is a phenomenal amount of money, which reflects the vast quantity and complexity of Federal records. The scale of digital acquisitions at a national repository is staggering. The UK National Archives Digital Archive acquired 7.8 gigabytes between June 2003 and May 2004. In the following three years, it expects to acquire over 10 terabytes, or over 43760% more material per annum.[76] Can institutions caring for the digital records of individuals ever hope to embark on such ambitious programmes, especially when we cannot, yet, provide our funders with realistic forecasts

of on-going costs? The Espida project at Glasgow University is currently investigating 'the relationships, roles and responsibilities, costs, benefits and risks inherent in institutional digital preservation'. The Glasgow team acknowledges that to date there has been little experience of implementing and assessing the costs and benefits of digital preservation to a specific community.[77]

Another key area of consideration relates to infrastructure - should all institutions be implementing solutions, or should there be centres of expertise?[78] Given the likely cost implications of ongoing commitments to digital preservation it would seem that national, or regional, centres of excellence are the way forward. Paradigm will investigate the benefits of collaborative as opposed to individual systems, which can treat these problems in a coherent and strategic manner and investigate how distributed modes of discovery and access might be used when the archives are opened. To date most of the UK research into digital preservation has come through the National Archives and HE/FE projects funded by bodies such as JISC. It seems likely that most universities will engage with digital preservation, if only for the more limited purpose of preserving their own digital research outputs. Interestingly TNA's Digital Preservation Department may see the development of 'off the peg' digital preservation packages developed with local authority record offices in mind as part as their wider remit to lead the UK archival profession.[79] Unless there is an initiative along these lines, it seems unlikely that local authority archives, and many specialist repositories, will have the resources, or expertise, to embark upon digital preservation.

How long will it take to develop a fully functional digital repository? [80] Arguably, it will never be entirely finished because technical development will continue to throw new issues our way. Certainly, we will need to develop preservation strategies for new file formats and evolve strategies for those formats that already exist in our archives. The architectures required for digital repositories will need to adapt and change to meet future developments. More frustratingly, as a team of researchers from Stanford University noted, 'The failure of a digital preservation system will become evident in finite time, but its success will forever remain unproven'.[81] As the core functions established in the OAIS model are perfected, it is likely that development will move to honing end-user searching and presentation systems. This aspect is not so urgent for digital private papers being collected now because they will be closed. The fact that they are closed is itself an issue. The hardest part may be convincing funding bodies and institutions of the need for extensive and ongoing funding for preserving digital materials that may not be generally accessible for decades.[82]

Should we be devising a new post-custodial model for personal papers in which individuals maintain their own digital records during their life-time? Guidance could be provided by heritage institutions until, as in the traditional scenario, the records are formally accessioned into the archive as the individual nears the end of their life, or reach the archive via the family once the individual has died. This would require the depositor to have an understanding of the issues of authenticity; a relatively high degree of IT nouse; and a commitment of time to, and enthusiasm for, disciplined record keeping that may not be realistic. Perhaps a simple and open format, such as the Open Document Format,[83] would meet the basic requirements for many of the record types produced by most authors or politicians. If such formats can preserve the appearance and content of a

typical office document, this would be a substantial part of the battle won. Complex file formats do not make up the majority of records and perhaps the key is to concentrate on a handful of popular formats. Widespread interest in the longevity of personal digital media might persuade software manufacturers, as in the case of Adobe, of the commercial possibilities of archival formats. However, there are inherent dangers in proprietary formats, from IPR issues to dependence on one commercial company, and therefore it seems likely that the most promising future for the preservation of personal material lies with the widespread adoption of open standards by commercial and non-commercial developers.

Will the ascendancy of the digital archive fragment the archive profession or indeed give rise to a new profession? Given the pace of technological change is it any wonder that many archivists are left feeling like 'Scribes in the age of Gutenberg.'[84] In future, will we see training courses developed specifically for digital archivists with a much greater emphasis on IT skills including basic programming and a good understanding of open source and digital repository software? There is also a case for rejecting the title of 'digital archivist' as the new profession is likely to cut across the sectoral boundaries which have traditionally divided the remits of IT professionals, museum curators, librarians, archivists and records managers. The digital world needs to utilise skills of all these people, and perhaps 'digital curator' would be a more useful term for those responsible for the management and long-term preservations of a wide range of digital records for the duration of the record cycle. Digital preservation borrows much from IT professionals and perhaps digital archivists are closer to the IT world than the archive profession. Recording and indexing 'born digital' material requires new skills which will inevitably lead to major changes in our approach to many archival functions, not least how we create catalogues. Will it be necessary to catalogue much below the collection level, particularly if the collection is open and the contents of the digital archive searchable by the user.[85] Indeed, given the vast amount of digital material, which is likely to be deposited in the near future, will we have time to catalogue below the collection level? Philosophical questions arise too. If a digital object must undergo repeated migrations as part of the ongoing preservation process the whole concept of the 'original' is lost. Each time a file is rendered, it is only a representation of the original.[86] Some of the tangible sense of history may be lost. A great poet may have authored the words you see on the screen, or printout, but there is no artefact to link the reader across time to the author. There is no digital equivalent to touching a piece of paper and knowing that a historical figure once held it too.

The authors hope that this paper will generate discussion. We have deliberately (and sometimes provocatively) raised questions and issues which the archival profession must address if we are to continue to effectively preserve the personal papers of individuals. It is important to remember that digital preservation is still in its infancy and, like the IT industry on which it depends, is rapidly evolving. It could well be that many of our musings on digital preservation, and its implications for the profession, turn out to be false starts. But this is no reason to procrastinate and avoid taking those initial steps. Our descendents might not be so easily persuaded that technology will find a way when faced with obsolete unrecoverable data and a historical record devoid of the personal.

[1] Many technical aspects of digital preservation have been omitted from this article for reasons of space and because we were just beginning our practical explorations in these areas at the time of writing. We hope to publish a second article towards the end of the Paradigm project which will offer critiques of the OAIS reference model, the METS and PREMIS metadata standards, the Fedora and DSpace digital repository software, as well as other tools, software and standards tested by the project team. A follow-up article would also give a detailed evaluation of the practical lessons learnt by the Paradigm project.

[2]For more information relating to this project see: Semple, Najla. 'Developing a Digital Preservation Strategy at Edinburgh University Library.' *Vine* 34 (2004): 33-37.

[3] CEDARS, or CURL Exemplars in Digital Archives, began in 1998 and ended in 2002. See http://www.leeds.ac.uk/cedars/ (accessed 6 February 2006).

[4] The 404 error is a standard response code generated by the Hyper Text Transfer Protocol (HTTP); it indicates that the web browser was able to communicate with the server, but the server either could not find the item requested, or was unwilling to fulfill the request. Not a good advert for digital asset management!

[5] De Rosa, Cathy, Lorcan Dempsey, and Alane Wilson. 2004. The 2003 OCLC Environmental Scan: Pattern Recognition. Copyright © 2004, OCLC Online Computer Library Center, Inc. OCLC Control Number: 53934212. Available at: http://www.oclc.org/membership/escan/toc.htm. Excerpt used with permission

[6] For further information regarding the programme and the projects funded see the homepage at http://www.jisc.ac.uk/index.cfm?name=programme_404 and Carpenter, Leona. 'Supporting Digital Preservation and Asset Management in Institutions.' *Ariadne* 43 (2005). See http://www.ariadne.ac.uk/issue43/carpenter/ (both accessed 6 February 2006).

[7] For more information on the requirements, functions and use of digital preservation in an institutional repository context see Wheatley, Paul. 'Institutional Repositories in the context of Digital Preservation.' *Technology Watch Report: Digital Preservation Coalition*, report 04-02 (2004).

[8] Digital Preservation Coalition http://www.dpconline.org/; ERPANET http://www.erpanet.org (both accessed 6 February 2006).

[9] DSpace project http://www.dspace.org; KB/IBM Long-Term Preservation Study http://www.kb.nl/hrd/dd/dd_onderzoek/dnep_ltp_study-en.html; CEDARS http://www.leeds.ac.uk/cedars/; CAMiLEON http://www.si.umich.edu/CAMILEON/; INTERPARES http://www.interpares.org; UKWAC http://www.webarchive.org.uk/ (all accessed 6 February 2006).

[10] The Digital Curation Centre's Associates Network, see http://www.dcc.ac.uk/associates (accessed 6 February 2006).

[11] CCSDS 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS).* Blue Book. Issue 1. January 2002. This Recommendation has been adopted as ISO 14721:2003 OAIS.

[12] One of the JISC's 4/04 projects has recently published an assessment of The National Archive's 'Digital Archive' and the National Digital Archive of Datasets (NDAD) compliance with the OAIS model. Beedham, Hilary et al. *Assessment of UKDA and TNA Compliance with OAIS and METS standards.* UK Data Archive, University of Essex, 2005. Available from http://www.jisc.ac.uk/index.cfm?name=project_oais (accessed 6 February 2006).

[13] Research Libraries Group. *Trusted digital repositories: Attributes and responsibilities*. An RLG-OCLC Report. (2002), available at http://www.rlg.org/longterm/repositories.pdf ; RLG and NARA. *An Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment* (August 2005), http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf (accessed 6 February 2006).

[14] DSpace http://www.dspace.org/ ; Fedora http://www.fedora.info/ (both accessed 6 February 2006). For an assessment of DSpace as an OAIS, see Tansley, Robert. Bass, Mick. Smith, MacKenzie. 'DSpace as an Open Archival Information System: Current Status and Future Directions.' *Lecture Notes in Computer Science* 2769 (2004): 446-60.

[15] The importance and complexity of the OAIS model is widely recognised and there are several introductions to the model available. One is Lavoie, Brian F. *The Open Archival Information System Model: an Introductory Guide.* Digital Preservation Coalition: Technology Watch Report (2004), available at http://www.dpconline.org/docs/lavoie_OAIS.pdf. (accessed 6 February 2006).

[16] OAIS, p. 1-1.

[17] The Preservation Metadata Implementation Strategies (PREMIS) Working Group. *Data Dictionary for Preservation Metadata.* RLG/OCLC, May 2005, available at http://www.oclc.org/research/projects/pmwg/premis-final.pdf . Maintenance of the PREMIS standard is being undertaken by the Network Development and MARC Standards Office of the Library of Congress, see http://www.loc.gov/standards/premis/ . (accessed 6 February 2006).

[18] OAIS, fig. 4-18, p. 4-37.

[19] For an introduction to the METS standard see Cantara, Linda. 'METS: The Metadata Encoding and Transmission Standard.' *Cataloging and Classification Quarterly* 40 (2005): 237-253.

[20] In a paper context we might check incoming archives for mould or pests. Ingest in a digital context involves similar actions (e.g. quarantine and checking for viruses and worms), but will also require additional processes, such as the validation of objects according to their format, the addition of technical metadata and even the transformation of objects into preferred formats (normalisation) to be retained alongside the original bitstreams.

[21] OAIS, fig. 4-1, p. 4-1.

[22] LOCKSS, see http://lockss.stanford.edu/ .(accessed 6 February 2006).

[23] Lynch, Clifford. 'Institutional Repositories: Essential infrastructure for Scholarship in the Digital Age.' *ARL Bimonthly Report* 226 (The Association of Research Libraries, February, 2003), available at http://www.arl.org/newsltr/226/ir.html (accessed 6 February 2006).

[24] EThOS is a consortium project funded by JISC, CURL and its partner institutions: the University of Glasgow, the British Library, Cranfield University, the National Library of Wales, the Robert Gordon University, SHERPA (a consortium led by the University of Nottingham), the University of Birmingham, the University of Edinburgh, the University of Southampton and the University of Warwick. See http://www.ethos.ac.uk for further details.(accessed 6 February 2006).

[25] There are currently five service providers: AHDS Archaeology; AHDS History; AHDS Visual Arts; AHDS Literature, Language and Linguistics; and AHDS Performing Arts. See http://www.ahds.ac.uk (accessed 6 February 2006).

[26] National Library of Australia's PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) http://pandora.nla.gov.au/; Internet Archive http://www.archive.org/; International Internet Preservation Consortium http://netpreserve.org/ (all accessed 6 February 2006).

[27] Sourceforge http://sourceforge.net/ (accessed 6 February 2006) .

[28] Cabinet Office. *Modernising government* London: Stationery Office, (Cm 4310), 1999, available at http://www.archive.official-documents.co.uk/document/cm43/4310/4310.htm (accessed 6 February 2006). TNA issued an invitation to tender to build a digital storage repository in 2002; Tessella was awarded the contract and designing, programming and testing took place 2002-3; see 'New Digital Archive at The National Archives' at http://www.nationalarchives.gov.uk/preservation/digitalarchive/pdf/project_background.pdf. For more information on TNA's approach to preservation, see Brown, Adrian. 'Automating Preservation: New Developments in the PRONOM service.' *RLG DigiNews* 9 (2005). To access material in TNA's digital archive, visit Electronic Records Online (ERO), http://www.nationalarchives.gov.uk/ero/ (all accessed 6 February 2006).

[29] Koninklijke Bibliotheek's e-Depot, http://www.kb.nl/dnp/e-depot/e-depot.html; National Archives of Australia Digital Preservation Project http://www.naa.gov.au/recordkeeping/preservation/digital/summary.html; NARA's Electronic Records Archive (ERA) programme at http://www.archives.gov/era/. (all accessed 6 February 2006).

[30] A notable exception is web-archiving.

[31] There are, of course, exceptions to this rule. Some organisations, including the Bodleian, have favoured more proactive collection development programmes (which include providing archival advice to potential depositors). We are also aware of institutions working with contemporary record creators who are particularly worried about the longevity of their email. Because much of this work requires absolute discretion, it is difficult to gage how much of it takes place.

[32] One example of recent media coverage, from a historical perspective, is Wojtas, Olga. 'Has the pen lost its might?' *The Times Higher Education Supplement*, 29 July 2005. The vulnerability of digital material was also highlighted by the media after many individuals and small businesses lost data when hurricane Katrina struck the Gulf Coast. For an interesting discussion on the vulnerability of personal records stored on a hard drive see Naughton, John. 'The platter that matters.' *The Observer,* 26 June 2005, page 6, retrieved from http://www.nla.gov.au/padi/qdigest/sep2005.html#2.9 (accessed 6 February 2006).

[33] For details of the Bodleian's modern political papers see http://www.bodley.ox.ac.uk/dept/scwmss/modpol/polpps.htm; for JRUL's political papers see http://rylibweb.man.ac.uk/data2/spcoll/. The Labour History Archive and Study Centre (LHASC) is based at the head office of the People's History Museum, for details of its collections see http://www.peopleshistorymuseum.org.uk (all accessed 6 February 2006).

[34]For more information see http://dspace.org and http://www.fedora.info (accessed 1 March 2006).

[35] The OAIS model provides guidance on preservation planning see Fig. 4-6, 'Functions of Preservation Planning' OAIS p. 61.

[36] These tools can only identify, validate and extract technical metadata from recognised and supported file formats. Extraction tools require detailed file format specifications, which can be retrieved from 'format registries'. There are several 'format registries' available, but the comprehensiveness and quality of their content varies. A leader in this field is PRONOM, a file format registry developed and maintained by the UK National Archives, see http://www.nationalarchives.gov.uk/pronom/. Other key players include The Global Digital Format Registry http://hul.harvard.edu/gdfr/ which has recently received a grant of $600,000 from the Andrew W Mellon Foundation,(both accessed 6 February 2006).

[37] Tools of interest include the National Library of New Zealand Metadata Extractor, which may be downloaded at http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction; the JHOVE tool http://hul.harvard.edu/jhove/, see ; and tools developed by the National Archives of Australia, see http://xena.sourceforge.net/ (all accessed 9 February 2006).

[38] A useful report which considers the two standards together is Lavoie, Brian and Gartner, Richard. *Preservation Metadata Technology Watch Report: Digital Preservation Coalition.* report 05-01 (2005).

[39] Adrian Cunningham, of the National Archives of Australia, has advocated that archivists working with personal papers should build close relationships with potential depositors and indeed seek to influence the way in which that person creates and manages their records. Cunningham first advocated this over a decade ago: 'Having secured an in-principle agreement for the eventual transfer of the person's records to the archives, the archivist will then need to build a lasting partnership with the donor whereby assistance is lent with the design of a recordkeeping system that satisfies predetermined standards and with the production of adequate support documentation'. See Cunningham, Adrian.'The archival management of personal records in electronic form: some suggestions.' *Archives & Manuscripts* 22 (1994):94-104, p 101.

[40] Jenkinson, Sir Hilary. "The English Archivist: a New Profession" in *Selected Writings of Sir Hilary Jenkinson* ed. Alan Sutton, Gloucester, 1980, p. 237. Jenkinson says: "Archives are the documents accumulated by a natural process in the course of the Conduct of Affairs of any kind, Public or Private, at any date; and preserved thereafter for Reference, in their own Custody, by the persons responsible for the affairs in question or their successors.".

[41] The project's Academic Advisory Board is a group of historians, political scientists and curators who offer advice on issues pertinent to the research communities, which will be using the digital materials collected by archivists as primary sources in the future. For more information, visit http://www.paradigm.ac.uk/about/aab (accessed 9 February 2006).

[42] Lynch, 'Institutional Repositories: Essential infrastructure for Scholarship in the Digital Age.'

[43] What survives to be archived, and indeed what is created in the first place, is down to what Sue McKemmish calls 'personal recordkeeping behaviours'. If archivists work with creators earlier, it is possible that our guidance will lead to more of the potential record types associated with personal archives being present in future collections, regardless of format. Will this mean that future collections of digital personal archives will be larger? Will they be artificial? See McKemmish, Sue. 'Evidence of Me.' *Archives and Manuscripts* 24 (1996): 28-45.

[44] For the full catalogue of the Morley Papers see http://www.bodley.ox.ac.uk/dept/scwmss/wmss/online/modern/morley/morley.html (accessed 7 February).

[45] For a collection level description of the Heffer Papers see http://www.archiveshub.ac.uk/news/0403eh.html (accessed 7 February).

[46] William Clark, radio interview in 1979 marking the 23rd anniversary of the Suez Canal Crisis (Bodleian Library, Oxford, MS. 145, f. 149), quoted by Helen Langley, 'Major Political Collections in the Bodleian Library, Oxford.' *Primary Sources and Original Works* 3 (1994): 93-112, p. 96.

[47] Dalgliesh, Paul. 'The Appraisal of Personal Records of Members of Parliament in Theory and Practice.' *Archives and Manuscripts:* 24 (1996): 86-101, p. 88. Other articles from this themed issue of *Archives and Manuscripts* on 'Personal Recordkeeping: Issues and Perspectives' are also worth consulting.

[48] Much of what Paradigm has accessioned was created in the last 5 years, some accessions include material created on the day of accession.

[49] These worries have prompted us to implement practical measures, such as the use of biometric technology to encrypt data in transit The project currently uses USB portable hard drives with fingerprint access control.
[50] The Paradigm deposit agreement can be seen at http://www.paradigm.ac.uk/workbook/accessioning/documentation/index.html (accessed 8 February 2006).
[51] The long-term survival of archival records created by small organisations, non mainstream community organisations, small businesses and pressure groups are also a matter of concern. This is particularly true of short-lived campaign groups such as those connected to the anti-globalisation movement. Not only are such records predominantly based on internet technologies but as they transcend national boundaries fall outside of national collecting remits.
[52] In fact, one of our politicians is digitising paper records as well as creating born-digital records. This digitisation is not simply the creation of basic digital surrogates, but includes the use of Optical Character Recognition technology to enable full-text searching.
[53] A copy of the survey document can be seen at http://www.paradigm.ac.uk/workbook/record-creators/surveying.html (accessed 1 March 2006).
[54] The project's *Workbook on Digital Private Papers* contains useful how-tos for these procedures. See http://www.paradigm.ac.uk/workbook (accessed 9 February 2006).
[55] It is worth noting that the accessions procedure can be very time-consuming, especially where large quantities of data must be copied and where the archivist must export email from software such as Microsoft's Outlook client.
[56] According to Michael Lesk of Bellacore; see Brand, S. and T. Sanders 'Escaping the Digital Dark Age.' *Library Journal* 124 (1999): 46-8, p. 47.
[57] Memories for Life is a Grand Challenge for Computing Science proposed by the UK Computing Research Committee, see http://www.memoriesforlife.org/. Quotation is taken from a 'Memory for Life' research paper, Fitzgibbon, Andrew and Reiter, Ehud. 'Memories for Life: Managing information over a human lifetime.' p. 2. See the PDF available at http://www.nesc.ac.uk/esi/events/Grand_Challenges/proposals/Memories.pdf. (all accessed 9 February 2006).
[58] Services might be free, but can they guarantee permanent access; will they protect your data in the event of a natural or man-made disaster? The small print is conspicuously absent in some cases.
[59] This number is unsurprising, given the ease of establishing and maintaining a blog. Services such as http://www.blogger.com enable users to create and host a weblog for free. See Technorati http://www.technorati.com/ for up-to-date statistics on the number of blogs; Google Blog Search is available at http://blogsearch.google.com/ (all accessed 9 February 2006).
[60]; Ourmedia http://www.ourmedia.org. (accessed 9 February 2006).
[61] For an excellent overview of personal digital collecting and the potential implications for heritage institutions see, Beagrie, Neil. 'Plenty of Room at the Bottom? Personal Digital Lirbaries and Collections.' *D-Lib Magazine*, vol. 11 No. 6 (June 2005), http://www.dlib.org/dlib/june05/beagrie/06beagrie.html (accessed 1 March 2006).
[62]Jenkinson, Sir Hilary. *Manual of Archive Administration*. London: Clarendon Press, 1922.
[63] Arguably a third strategy is 'digital archeology' which involves directing large amounts of money and highly trained IT specialists to recover obsolete data. The cost implications for this strategy make it an unattractive option, and the decay of the manufacturing facilities which produced old parts renders it unviable in the long-term. Nevertheless, this approach will remain part of the digital archivist's toolkit as long as we need to 'rescue' high value collections. See report by Ross, Seamus and Gow, Ann. 'Digital Archaeology: Rescuing Neglected and Damaged Data Resources.' *JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials*. (1999):1-94.
[64] Further information regarding the digital preservation strategy adopted at the National Archives of Australia is available at http://www.naa.gov.au/recordkeeping/preservation/digital/summary.html (accessed 9 February 2006).
[65] For more information on Emulation see the CAMiLEON project, http://www.si.umich.edu/CAMILEON/ (accessed 1 March 2006) and Granger, Stewart. 'Emulation as a Digital Preservation Strategy', *D-Lib magazine*, October 2000.
[66] Let's take a website as an example: the user's experience will depend on the software they are using (e.g. web browser and operating system) as well as the hardware they are using (e.g. size of screen, speed of processor).
[67] For a good introduction to the main issues when preserving digital material see chapter two in Jones, Maggie and Beagrie, Neil. 'Digital Preservation.' *Preservation Management of Digital Materials: A Handbook,* British Library 2001 (reprinted 2003). A regularly updated version of this handbook is also available on-line at http://www.dpconline.org/graphics/handbook/ (accessed 1 March 2006) .
[68] Oltmans, Erik and Kol, Nanda. 'A Comparison between Migration and Emulation in Terms of Cost' *RLG DigiNews*, vol. 9, no. 2. Available from http://www.rlg.org/en/page.php?Page_ID=20571&Printable=1&Article_ID=1714. The results of the Life project, which aims to cost different elements of the digital curation lifecycle, may also be worth consulting; see http://www.ucl.ac.uk/ls/lifeproject/. (both accessed 9 February 2006).
[69] For an excellent overview of open source software, especially as it related to digital curation, see McHugh, Andrew. 'Open source for Digital Curation', *DCC Digital Curation Manual*, July 2005. See http://www.dcc.ac.uk/resource/curation-manual/chapters/open-source/ (accessed 1 March 2006).
[70] Peeling, Dr Nic and Satchell, Dr Julian. *Analysis of the Impact of Open Source Software*. Report published online by QINETIQ Ltd 2001. Available at http://www.govtalk.gov.uk/documents/QinetiQ_OSS_rep.pdf (accessed 1 March 2006).
[71] The establishment of JISC's Open Source Software Advisory Service (OSSWatch) is evidence of this. See http://www.oss-watch.ac.uk (accessed 9 February 2006).
[72] Numerous technology companies are now involved in open source software developments. For example, Sun http://www.sunsource.net/; IBM http://www-128.ibm.com/developerworks/opensource; Google http://code.google.com/; Hewlett Packard http://opensource.hp.com; Novell http://developer.novell.com/opensource/index.html?sourceidint=hp_developers_novell-opensource; and My SQL http://www.mysql.com (all accessed 1 March 2006).
[73] The first part of the international standard about PDF/A format was officially published by ISO on 28th September 2005, under reference ISO 19005-1 'Electronic document file format for long-term preservation – Use of PDF 1.4 (PDF/A-1)'. PDF/A-1 format conforms to PDF 1.4 format but does not use all features of PDF 1.4, in order to allow the better preservation and display of documents. It is applicable to documents containing combinations of character, raster and vector data. Sound and video are not permitted. For more information on PDF/A see http://www.aiim.org/documents/standards/PDFreference.pdf (accessed 1 March 2006).
[74] Talbot, D. 'The Fading Memory of the State' *Technology Review* 108 (2005): 44-9.

[75] NARA press release, September 8th, 2005 which states that the new Electronic Records Archive system for NARA 'will capture electronic information – regardless of its format – save it permanently, and make it accessible on whatever future hardware or software is currently in use.'
[76] *The National Council on Archives report*, 'Your Data at Risk: Why you should be worried about preserving electronic records', September 2005: 6.
[77] For more information see http://www.gla.ac.uk/espida/ (accessed 1 March 2006).
[78] One idea might be a distributed digital repository prototype for personal papers be they those of politicians, authors, scientists or musicians.
[79] Brown, Adrian, 'Preserving the digital heritage: building a digital archive for UK Government records'. *Online Information 2003 Proceedings:* 65-68. Available at http://www.nationalarchives.gov.uk/preservation/digitalarchive/pdf/brown.pdf (accessed 1 March 2006). In the concluding paragraph of this article Adrian Brown discusses how in future the TNA will look at providing guidelines on preserving digital records at a local level.
[80] One good example of a relatively mature digital archiving infrastructure is the Californian Digital Library, which began life as a single post with some money for travel expenses. See Caplan, Priscilla. 'Building a Digital Preservation Archive: Tales from the front.' *Vine* 1 (2004): 38-42.
[81] Rosenthal, David S. H. et al 'Requirements for Digital Preservation Systems: A Bottom-Up Approach.' (2005): 11 http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cs/0509018 (accessed 1 March 2006).
[82] It may be possible to open some series of records, especially those falling outside the remit of the Data Protection Act, sooner than others. However this will require detailed negotiation with depositors, many of whom will still be actively engaged in their working lives and have limited time for protracted negotiations. Some depositors may prefer to issue blanket restrictions.
[83] The Open Document Format, is an open, XML-based, format for office documents. The standard was created, and is maintained by, the Organization for the Advancement of Structured Information Standards (OASIS). It has also been submitted to the International Standards Organization for approval as an ISO standard. See http://www.oasis-open.org/committees/office/faq.php (accessed 1 March 2006).
[84]Coined by John Hodgson, Keeper of Manuscripts and Archives at the John Rylands University Library, The University of Manchester.
[85]At a workshop held by historians over 12 years ago it was noted that the traditional worlds of archives, libraries and museums are challenged by digital media, 'Simple notions such as document, sequence and provenance are already gravely compromised', p 309. Morris, R J. 'Electronic documents and the history of the late 20th century: Black holes or warehouses-What do historians really want?' In *Electronic Information Resources and Historians: European Perspectives*, edited by Seamus Ross, and Edward Higgs, London, 1993. Proceedings of a workshop held at The British Academy on 25 & 26 June 1993.
[86] The Preservation Metadata Implementation Strategies (PREMIS) Working Group. *Data Dictionary for Preservation Metadata*, RLG/OCLC, section 1-10 'It is not possible to change a file (or bitstream or representation); one can only create a new file (or bitstream or representation) that is related to the source Object.'