PARADIGM

A practical approach to the preservation of personal digital archives

Susan Thomas
Bodleian Library
March 2007
v 1.0

This is the final report of the Paradigm project, which was funded by the Joint Information Systems Committee (JISC) under its Supporting Digital Preservation and Asset Management in Institutions 4/04 Programme and the University of Oxford's Research Development Fund (now the John Fell OUP Research Fund).

Project website: http://www.paradigm.ac.uk/

JISC Programme website: http://www.jisc.ac.uk/index.cfm?name=programme_404

Project contact: Susan Thomas (Project Manager),
susan.thomas@bodley.ox.ac.uk,
01865 2838321

Table of Contents

1 Executive Summary

1.1 Introduction

The Paradigm project represents a pioneering effort to begin developing practical solutions to the challenges facing research libraries that wish to continue enriching their archival collections in our increasingly digital times. Paradigm deals exclusively with the archives of individuals – personal archives – and is among the first significant undertakings of its kind. Personal archives are an important part of our cultural memory, but inexperience in curating their born-digital counterparts puts the survival of today's personal histories at risk. The diversity and volatility of digital technology far exceeds that of any medium that creators, archivists and researchers have previously worked with. While materials created digitally are forming an increasing part of the holdings of university special collections departments, most of these institutions have not yet established means to collect, process, preserve or provide access to the digital leavings of significant individuals. The challenges posed by personal digital archives, notably the need for ongoing preservation awareness and activity to begin much earlier in the lifecycle, invite the development of alternative strategies to accomplish these activities. Future researchers will not benefit from access to the unpublished materials accumulated and created by our significant contemporaries unless practical changes can be made soon to the ways in which universities deal with archival collecting and collections.

With the Paradigm project the Bodleian Library, University of Oxford, and John Rylands University Library, University of Manchester, have taken their first practical steps in developing digital curation and preservation capabilities for archival materials. Paradigm has investigated what it means to curate and preserve mixed media, or hybrid, archives containing born digital and more traditional materials. It has engaged with personal archives and their creators and the researchers who will use such archives in the future. Paradigm has tested various tools, techniques and standards that have emerged from the digital curation and preservation communities alongside those that have been used by archivists of old. It is hoped that the project can act as a useful exemplar for individuals and institutions interested in the preservation of personal digital archives and to this end the online Workbook produced by the project documents some of the learning, techniques and best practices developed by the project.[1] Although the project has focused on personal archives, many of the lessons learned will be applicable to the collection and management of digital materials that share some or all of the characteristics of a personal archive:

- materials created by individuals using a variety of standards, tools, and technologies over which the archivist has little or no direct influence;
- materials whose authenticity must be preserved;
- materials with important contextual relationships;
- materials in which several parties have privacy or intellectual property rights;
- materials for which traditional approaches to appraisal are inappropriate;
- materials which will be preserved indefinitely.

The project should also be of interest to those working with digital curation and preservation standards, tools and repositories in other areas.

This report provides an overview of the Paradigm project's findings and recommendations.

Finally, the project staff wish to thank the Oxford Research Development Fund and the JISC for

[1] Paradigm Workbook http://www.paradigm.ac.uk/workbook

their generous funding of this project.

1.2 Findings in brief

Working with depositors and their materials, and talking to potential researchers,[2] has allowed the project to identify similarities and differences between working with digital and physical archives, and to develop a better understanding of the practical requirements for processing born digital archives throughout their lifecycle. Some ideas for best practices have been published in the Workbook and areas which require further research and development for practical implementation have been identified.

Among the project's findings are that:

- the storage of personal digital archives is distributed across local devices, storage media, networked servers and web services. In many ways their component parts are less obvious than those of physical archives and may be overlooked by a depositor or accessioning archivist;
- individuals often fail to see contemporary materials as archives;
- digital preservation awareness amongst creators is low, even relating to preservation for their own short- to medium-term needs;
- individuals seem more concerned about the security and privacy of digital materials as they are viewed as being more easily copied and distributed;
- there may be duplication between digital and paper materials in a single archive;
- archivists may need training in extracting certain kinds of digital archive, such as email. More how-to guides, such as those in the Paradigm Workbook, would be helpful;
- research libraries collecting personal digital archives may need to employ a variety of collection strategies to accommodate the personalities of creators and provide the best possibility of strong digital archives collections. This should embrace older digital technologies through to contemporary ones;
- none of the file formats accessioned from contemporary politicians (ranging from days old to five years old) were inaccessible using contemporary computing environments, although some required the use of proprietary software;
- file format registries do not contain information on some important contemporary and legacy file formats found in personal digital archives;
- there is no registry of file format migration tools that could help archivists locate tools known to produce good quality file format migrations;
- preservation requires the recording and association of a variety of metadata that is unfamiliar to traditional archivists. Digital archivists must get to grips with a variety of metadata content and mark-up standards;
- digital archivists need accessible information about the technical attributes of file types and file formats;
- metadata extraction tools are varied in the ease of their installation and use, and the metadata that they generate is often tool-specific rather than standards-based. The output of such tools must be assembled into metadata packages, often requiring supplementary archivist-supplied metadata. Metadata generation must be greatly simplified if the preservation of personal digital archives is to to be workable;

[2] More details of Paradigm's Academic Advisory Board are available at http://www.paradigm.ac.uk/about/aab/index.html

- open source digital repositories, such as DSpace and Fedora,[3] provide a useful managed environment for the preservation of digital materials, but configurations and use cases for the needs of archivists require development. Some useful preservation monitoring facilities are planned for these repositories by their respective development communities;
- Paradigm finds Fedora to be the better option for archivists, for several reasons including its extensibility and its ability to support complex objects and their hierarchical relationships;
- digital forensics and data recovery software are useful tools for capturing and exploring digital archives;
- means of persistently identifying digital documents and their successor migrations must be developed. Researchers must understand what version of an archive they are using and how it should be cited in publications;
- changes to Freedom of Information legislation that allow private archives donated to public institutions to be excluded from the provisions of the legislation for a number of years would aid collection development, particularly collection development that aims to acquire archival materials sooner after creation;
- changes to copyright legislation that expressly permit archives and libraries to undertake preservation actions, such as the creation of redundant copies and format shifting are needed;
- it is difficult for researchers to appreciate what working with digital archival materials will mean until they experience it, but research libraries should endeavour to involve the research community in digital archives work to promote awareness of the changing nature of archival materials.

1.3 Summary outcomes

The outcomes of the project include:

- better appreciation of the scope and content of the materials that we might expect older, contemporary and even future personal digital archives to contain;
- guidelines for creators of personal digital archives, to assist creators in thinking about safeguarding their digital materials;[4]
- participating institutions have a platform for developing the capacity to continue selecting, taking-in, managing, cataloguing, and making personal archives available to researchers;
- the application of digital preservation tools and standards to traditional archival workflows has allowed archivists to practice key events in the lifecycle of digital records, such as accession and ingest, and to develop procedures accordingly;
- testing of digital repository, other digital curation tools and metadata standards has provided an understanding of their strengths and weaknesses;
- participating institutions have a better understanding of the skills, infrastructure and processes needed to administer and preserve digital archives. This has led the project partners to join with the Wellcome Library in undertaking the Cairo software development project for ingesting archives into digital repositories;[5]
- by working with real creators of personal archives and their materials, we have developed a better understanding of the issues that are important to the individuals whose papers we

[3] Dspace website http://www.dspace.org; Fedora website http://www.fedora.info.

[4] Paradigm's Guidance for Creators of Personal Digital Archives http://www.paradigm.ac.uk/guidanceforcreators/index.html

[5] Cairo project website http://cairo.paradigm.ac.uk

collect, and how different collecting strategies might affect both our relationship with them and the resulting archive;

- a prototype digital archive system, using the Fedora digital repository software, has been established at the Bodleian. This does not yet possess all the desirable attributes of a trusted digital repository but represents a managed environment that is a considerable improvement on the storage of digital materials in a stack environment.

1.4 Summary conclusions

Conclusions drawn from Paradigm's experiences comprise the following:

- providing creators with access to digital curation expertise may help in increasing the likelihood of digital archive survival;
- existing skills of archival professionals will be critical to the management of personal digital archives, but archivists dealing with digital materials will need to extend their skills portfolio to embrace the requirements of digital materials;
- existing skills of system administrators and software developers are critical to the management of personal digital archives, but must be extended to understand archival requirements;
- institutions wishing to engage with digital preservation need access to expertise, whether this means giving existing staff new responsibilities, third party expertise or dedicated in-house expertise. No end-to-end, user-friendly, systems for ingest, administration, preservation and access for digital archives exist to hide the complexities of digital curation and preservation activities. The current generation of digital curators must therefore be conversant with a far greater, and constantly evolving, range of tools and standards than archivists caring for traditional materials. While some aspects of processing digital archives can be undertaken by traditional archivists with minimal training, other activities would require significant training that is unrealistic while paper materials still form the more significant quantity of materials.
- incoming digital materials are of a lesser quantity than traditional materials, and while there is a scarcity of staff conversant with digital curation and preservation, it is understandable that digital curators will be in the minority. This leads to a lack of support and understanding of the role of a digital curator within institutions that can be alleviated, to some extent, by collaborative working across institutions and by promoting awareness of the high level issues. The balance of digital and traditional skills in archives may evolve over time;
- institutions need an infrastructure supported by Information Technology professionals so that archival professionals can concentrate on the aspects of collection, management and access of materials.
- there will be no final solution to the preservation of personal digital archives. Constant re-evaluation will be required, as computer hardware, software, formats and usage evolves. The extent of learning and change required should be less once high level policy and procedure is in place, but there will always be a need for a “Technology Watch” service that monitors industry developments and individual creators' practices so that preservation capability is extended to cater for changes in the digital landscape;
- digital preservation need not be all or nothing. Small initial steps can be taken to start learning about and working with digital archives that do not entail full blown implementations of sophisticated digital repository systems. It is important not to delay much needed activity by waiting for a perfect system that will never appear. Experience will help defined the systems needed.

1.5 Recommendations

There follows a list of the recommendations presented in the body of the report.

Recommendation 1: research libraries should involve researchers in formulating collection development policy and strategy and in identifying potential collections of interest. *See p. 15*

Recommendation 2: more case studies that test different approaches to collection development are needed; from accessioning old media and devices to accessioning contemporary files. *See p. 15*

Recommendation 3: case studies of working with personal archives and their creators in other domains, such as science and literature, are required in order to learn about the social, legal and technical issues pertinent to the whole range of creators of personal digital archives commonly collected by university special collections. *See p. 15*

Recommendation 4: an analysis of the impact of the digital medium on the trade in archives and manuscripts would be of use to research libraries. *See p. 16*

Recommendation 5: digital archivists should maintain an awareness of the range of potential locations and mediums used for the storage of digital archives by their creators, and talk to creators about such materials to ensure that they are not overlooked. *See p. 17*

Recommendation 6: research libraries should use record surveying techniques when talking to creators and surveying materials; screenshots and directory trees can be very helpful. This information can be useful for a number of digital, and general, archive administration activities. *See p. 18*

Recommendation 7: digital archivists should talk to researchers about personal digital archives to inform selection decisions and to increase awareness of the changing shape of personal archives. *See p. 19*

Recommendation 8: research libraries should become familiar with means of proving authenticity in the digital environment, such as checksums and digital signatures. *See p. 19*

Recommendation 9: research libraries should become familiar with the extraction of digital archives from popular software, devices and web services. *See p. 20*

Recommendation 10: research libraries should publish how-to guides for accessioning common kinds of personal digital archives for future reference and the benefit of others. *See p. 20*

Recommendation 11: research libraries should explore the potential of digital forensics hardware and software for survey and extraction of personal digital archives at the creators' premises and at the repository; a comparison of open source and proprietary solutions would be useful. *See p. 20*

Recommendation 12: research libraries should identify trigger points that may bring about the destruction of significant personal archives and aim to accession such materials if appropriate. *See p. 21*

Recommendation 13: research libraries should aim to document third-party interests in archival collections they acquire so that these can be managed appropriately. *See p. 21*

Recommendation 14: research libraries should use records surveying techniques to ensure that they are well-prepared to make accessions. *See p. 21*

Recommendation 15: research libraries should be able to estimate the time needed to undertake an accession. *See p. 21*

Recommendation 16: research libraries should develop well-defined, but flexible, approaches to collection development and flexible accessioning procedures. *See p. 22*

Recommendation 17: research libraries need to make a commitment to personal digital archives, moving this work from a project to a permanent footing. Contact with depositors whom research libraries wish to work with in the long-term should be made by permanent staff and form an integral part of their daily work. *See p. 22*

Recommendation 18: research libraries should be able to reassure creators that material subject to embargo is held securely. *See p. 22*

Recommendation 19: in order to encourage creators to deposit archives with research libraries sooner after their creation, research libraries should be able to reassure creators that private materials given to them are not subject to Freedom of Information provisions while the creator has placed an embargo on their access. *See p. 22*

Recommendation 20: research libraries should offer generic or specific advice to the creators whose personal archives they are accessioning and/or aim to keep in touch with potential donors to monitor the technologies that will be coming to the repository in due course. *See p. 22*

Recommendation 21: more practical research into post-custodial collection development, where creators care for their own archives using guidance and systems provided by a research library, would be useful, though this may need to be undertaken over several years. *See p. 23*

Recommendation 22: special interest groups (e.g. archivists working with literary, scientific or political personal archives) should raise awareness of digital preservation issues particular to specific domains by targeting their advice to those audiences. *See p. 23*

Recommendation 23: research libraries should keep a watching brief on developments in webarchiving initiatives and technologies. *See p. 24*

Recommendation 24: if research libraries intend to use third parties to process obsolete digital archives they should include permission to do so in deposit agreements. *See p. 24*

Recommendation 25: the preservation community should lobby for changes that would see vendors rescind commercial interests in obsolete software, or change the law so that software licences contain an expiry date after which their use is not constrained by the terms of the original licence. *See p. 24*

Recommendation 26: websites that record information about extracting data from past and present personal computing environments should be submitted to UKWAC or the Internet Archive for archiving by research libraries. *See p. 25*

Recommendation 27: research libraries should create how-tos for the extraction and migration of common obsolete digital media and file formats. *See p. 25*

Recommendation 28: research libraries should be able to access preserved and working examples of the hardware and software needed to extract data from older computing environments. *See p. 25*

Recommendation 29: research libraries should promote the use of open standards that pose fewer problems for the preservation community and others. *See p. 25*

Recommendation 30: research libraries should investigate how best to provide secure access to archival storage for archival staff over a network. *See p. 26*

Recommendation 31: research libraries should be aware of the incidental copies of digital materials made by computers and how digital material is securely deleted. *See p. 26*

Recommendation 32: research libraries must develop business rules for the treatment of data types and data formats so that digital preservation can be as consistent, predictable and economical as possible. *See p. 26*

Recommendation 33: a central repository providing information, and perhaps access to tools, needed to identify good quality migration pathways would be helpful in devising business rules for the treatment of formats submitted to research libraries. *See p. 26*

Recommendation 34: research libraries should contribute information to file format registries so that their content can be developed to include a wider range of the formats typically found in personal digital archives. *See p. 29*

Recommendation 35: the digital preservation community should raise awareness of file format registries with open source and proprietary software developers. *See p. 29*

Recommendation 36: research libraries should monitor trends of functionality and use in the arrangement and use of personal digital materials. *See p. 30*

Recommendation 37: research libraries should encourage creators to organise materials rather than rely on searching technologies. *See p. 30*

Recommendation 38: research libraries should aim to record PREMIS metadata using the PREMIS XML schemas within METS files. This should increase the understandability of the metadata and facilitate interoperability. *See p. 31*

Recommendation 39: more accessible background information for technical metadata standards is needed by research libraries. *See p. 31*

Recommendation 40: research libraries need a single user friendly interface that combines the functionality of a number of metadata extraction tools to create METS-based Archival Information Packages. *See p. 32*

Recommendation 41: research libraries should develop a model for persistently identifying the component of personal digital archives carefully and commit to the maintenance of the system chosen. *See p. 32*

Recommendation 42: research libraries should monitor trends in the growth of personal digital archives to plan for storage needs. Growth includes incoming collections and their metadata, increasing metadata for existing archives and extra files created by format shifting, etc. *See p. 33*

Recommendation 43: research libraries should develop data models, and interfaces based on those models, for using Fedora as a preservation repository for personal digital archives. Fedora for archivists out-of-the-box would make its use a possibility at institutions with less technical support which might otherwise have to develop divergent local custom solutions. *See p. 34*

Recommendation 44: research libraries should perform local technology watch based on a knowledge of their holdings, and remain aware of technology watch services available in the digital preservation community and of digital curation and preservation trends. *See p. 34*

Recommendation 45: research libraries should engage with the research community to promote a wider understanding of digital preservation issues and their impact on the primary sources that will be available in future years. *See p. 36*

Recommendation 46: research libraries should investigate how access to personal digital archives will be delivered. *See p. 36*

Recommendation 47: research libraries should seek legal advice about the provision of digital materials both in a stand-alone reading room environment and over networks, and about the supply of copies for research use. *See p. 36*

Recommendation 48: research libraries should aim to progressively build experience in working with digital archives through direct engagement with creators and their personal archives. *See p. 37*

Recommendation 49: research libraries should take advantage of the offerings of organisations such as the DPC and DCC. *See p. 37*

Recommendation 50: graduate traineeships for new entrants into the archival profession, and archival courses and modules that allow students to specialise in born digital archives should be created. *See p. 37*

2 Introduction

Archives are materials selected for permanent preservation because they are perceived to have enduring value; they are an important part of academic and cultural life. Personal archives provide a personalised lens through which researchers can view the subject of their inquiry; such personal perspectives are often easily translated to wider society. Personal archives of significant individuals, such as politicians, writers and scientists have long formed a valuable part of the holdings of research libraries like the Bodleian and the Rylands.

Both libraries are expert curators of these materials in analogue forms, but the long-term survival of personal archives which document our own times is threatened, as more and more of the material we have come to associate with such archives is born digital. Digital materials present additional and ongoing preservation challenges which are well documented in the growing literature on 'digital curation' and 'digital preservation', but not generally well-understood by those creating the archives that researchers may wish to use in future, by the research libraries responsible for acquiring, preserving and providing access to such materials, or by the researchers who will be the users of these archives in years to come.

The challenges posed by the preservation of personal digital archives are distributed amongst these three groups – the creators, the research libraries and the researchers. Some of the technical challenges are common to the preservation of other kinds of archival materials, which allows us to borrow from work undertaken in these areas; the organisational challenges are more particular to institutions that collect archives, especially personal archives, from multiple external sources.

This is a period of transition for research libraries. On the one hand, there is no shortage of important existing and incoming traditional archival material demanding attention; on the other, there is an urgent need to engage with digital materials that have very different curatorial and preservation requirements. Increasingly these traditional and digital materials are combined in hybrid personal archives. Resources for traditional archival work are as necessary as ever, so the potential for redeploying resources to develop capacity for digital archives is limited.

Staff with the confidence, interest, knowledge and experience to work with digital archives are a rare breed; attracting existing archivists to work with digital materials can be difficult because digital files lack the allure of handwritten or annotated manuscripts and because working with digital materials requires much new learning of an unfamiliar terrain. Nevertheless, many of the challenges raised by the shift to digital media are well known and have long been discussed by the archival community. Intervening earlier in the lifecycle has been an important part of the solution to dealing with digital records for organisational records managers and for digital preservation more broadly. The idea of applying such techniques to the realm of personal archives is nothing new,[6] but has not been so well explored in practical terms, or at least this exploration has not been well documented for the benefit of the larger archival community.[7]

One potential benefit of an early-intervention approach is that research libraries can inform creators of digital preservation issues, thereby increasing the likelihood of survival for significant personal archives; it also allows archivists to work with digital materials that are contemporary, familiar and healthy, rather than obscure, corrupt or obsolete. One of the key questions Paradigm set out to answer is to what extent this kind of early-intervention approach can work in the case of personal archive collections. The other was how traditional archival workflows, tools

[6] Adrian Cunningham's articles on this matter in the 1990s are useful. See, for example, Adrian Cunningham, 'The Archival Management of Personal Records in Electronic Form: Some Suggestions', *Archives and Manuscripts* 22:94 (May 1994).

[7] Lucie Paquet's writings on experiences with personal digital archives at the National Archives of Canada are a notable exception. See 'Appraisal, Acquisition, and Control of Personal Electronic Records: From Myth to Reality' *Archives and Manuscripts* 28:2 (November 2000).

and standards used with personal archives might interface with those designed for the curation and preservation of digital materials.

2.1 Project scope

Paradigm was designed to be a very practical two-year pathfinder project. By piloting approaches, standards and tools with contemporary personal archive creators, personal archives and researchers it was felt that the project could achieve an accurate understanding of the issues and available tools that would provide a basis for realistic solutions.

2.1.1 Creators and archival materials

While the Bodleian and Rylands possess archival collections of immense breadth and depth,[8] Paradigm's focus was restricted to the personal archives of working politicians to keep the project more or less manageable. Working with contemporary politicians targeted an area of importance to both institutions and raised the profile of digital preservation issues with participating politicians. The project scope was slightly extended to include a small website harvesting pilot around the 2005 general election, and to work with older digital materials deposited at the Bodleian Library as part of the archive of former Cabinet Minister Barbara Castle. These additional case studies were useful comparisons to the project's primary case study.

The nature of the case studies piloted by Paradigm can be summarised as follows:

CASE STUDY NO.		CASE STUDY DESCRIPTION	COLLECTION DEVELOPMENT APPROACH
1		**Develop relationships with a variety of working politicians creating personal archives and acquire sample collections for testing digital curation and preservation techniques.**	
	1a	**Records survey and three digital accessions from Westminster office.**	**Snapshot**
	1b	**Records survey of materials at Charity, Westminster and European Parliament offices. Two paper accessions.**	**Postcustodial**
	1c	**Records survey at constituency office. Politician dropped out after unseating in 2005 General Election.**	
	1d	**Records survey and hybrid accession at Westminster office.**	**Transfer via retired hardware or media**
	1e	**Records survey and accession at constituency office.**	**Snapshot**
	1f	**Records survey and accessions at constituency and Westminster office.**	**Snapshot**
2		**Web archiving pilot to harvest select personal websites and blogs during the 2005 general election.**	**(Remote) snapshot**
3		**Digital archaeology pilot using two IBM PCs, Amstrad 3" disks and 3.5" disks that are part of the archive of Barbara Castle, Baroness Castle of Blackburn.**	**Traditional**

[8] Further information on the archival collections of both institutions is available at http://www.bodley.ox.ac.uk/dept/scwmss/wmss/wmss.htm and http://rylibweb.man.ac.uk/data2/spcoll/

2.1.2 Workflows

The focus was on managing and preserving hybrid personal archives throughout their lifecycle and fusing archival and digital curation/preservation tools and techniques to accomplish this. The project made extensive use of standards designed for the management of digital materials. The central standard underpinning much work in this area is the Reference Model for an Open Archival Information System (approved as international standard ISO 14721 in 2002).[9] This standard has been invaluable in connecting the terminology of archival and technical staff and Paradigm used this in combination with traditional archival workflows to develop a lifecycle model for hybrid personal archives and a prototype digital archiving system. This will be explored in the body of the report.

2.1.3 Tools

Paradigm was to test tools required for the lifecycle management of personal digital archives, such as metadata extraction tools. It also explored the potential of two digital repository systems – DSpace and Fedora - for preserving personal digital archives.

2.2 Aims and objectives

Paradigm's ambition was to start developing a realistic approach to curating personal digital archives that involved creators, curators and researchers in order to give research libraries the confidence to bring born-digital materials into their collections. It was conceived to address the issues associated with personal digital archives in the context of existing collecting and curatorial responsibilities, acknowledging that digital materials are currently the smaller component of incoming personal archives and form a small, but growing, part of the collections managed by research libraries. Broadly speaking, Paradigm's aims and objectives were:

- to provide the participating institutions with hands on experience of curating personal digital archives;
- to start developing the skills, experience and infrastructure to manage digital archive materials at the participating institutions;
- to pilot a system for selecting, taking-in, managing, cataloguing, and providing long-term access to digital personal material that could serve as a platform for future work at both institutions;
- to compare and contrast archival operations in the digital and traditional archival environments;
- to test metadata standards, tools and digital repository systems;
- to investigate the impact of intervening earlier in the lifecycle on the curation and preservation of personal digital archives by collecting personal archives from working politicians;
- to act as an exemplar for those beginning to work with personal digital archives by documenting the digital preservation knowledge acquired by the project team, and the application of that knowledge in team's work with politicians and their personal archives;
- to consult with researchers about aspects of collecting and curating born-digital personal archives that impact on the re-use of such materials;
- to produce an online Workbook[10] that details the project's experiences and provides

[9] Reference model for an Open Archival Information System: http://public.ccsds.org/publications/archive/650x0b1.pdf

[10] Paradigm Workbook http://www.paradigm.ac.uk/workbook

workable 'best-practice' guidelines based on these for the benefit of other archival institutions;

- to use the initial outcomes of the project as a catalyst for ongoing activity in this area.

2.3 Project methodology

From 2005 to 2007, Paradigm has explored many facets of the problem of preserving personal digital archives. The project's methodology has involved familiarisation with existing and emerging literature and project work and deriving lessons that can be applied alongside traditional archival principles and practice to the particular context of an organisation working with the archives of individuals. It has been an interdisciplinary project, involving all relevant parties at appropriate points in the lifecycle: creators of archival materials, archivists responsible for their care, IT professionals to support the building and maintenance of an infrastructure, researchers who might use them in the future and managers who need to understand the implications of the digital environment on the organisation. This has helped to establish what the obstacles to working with digital archives are so that workable solutions that instil confidence in all parties can be found. Because Paradigm was designed to give participating institutions a flavour of what it means to work with born digital archives, its approach has been very practical in nature:

- to unite the requisite archival and technical skills in a single interdisciplinary team;
- to survey and accession sample hybrid personal archives (i.e. personal archives with digital and physical components) from a representative sample of working Conservative, Labour and Liberal Democrat politicians;
- to assess the impact of the hybrid environment on the acquisition, processing, preservation and distribution of personal archives and to pilot workflows for processing and preserving hybrid archives that combine traditional archival theory and practice with emerging digital curation theories and practices, such as the OAIS model;
- to investigate digital repository software by developing a prototype repository for archival storage and preservation;
- to use the digital components of the politicians' exemplar personal archive collections to test tools and metadata standards applicable to the processing and preservation of the born digital archives;
- to develop an understanding of the legal and social issues around born digital personal archives;
- to seek input from an advisory board of academics on matters relating to the collection, processing and re-use of the materials relevant to researchers;
- to communicate lessons and best practice to the wider community.

2.4 Outcomes

At the partner institutions, the principal outcomes of the project include:

- greater knowledge of digital curation issues;
- improved understanding of what tools and techniques are available to work with digital archives now and what is forthcoming;
- improved support for digital archive encounters made outside of the project's framework in the course of the normal activities of the Special Collections departments;
- changes to existing institutional policies to take into account the acquisition and

preservation of digital archives;

- prototype Fedora-based digital repository for archival storage of digital archives and metadata at the Bodleian;
- findings around metadata, ingest and preservation monitoring require further research and development to produce workable solutions; to this end the Bodleian and the JRUL recently joined with the Wellcome Library in the Cairo project,[11] which has been funded by JISC under the 'Tools and Innovations' strand of the *Repositories and Preservation Programme;*
- desire to build on the work of Paradigm.

Wider outcomes include:

- greater awareness, among both the archival and the digital preservation communities, of personal digital archives and the problems and potential solutions for their ongoing preservation;
- availability of project documentation for others interested in the preservation of personal digital archives; certain outputs of the project are easy to implement, such as surveying techniques and template deposit agreements; others, such as repository implementation and commitment to collecting and preserving digital archives require policy decisions and ongoing commitments to specialist staffing and technologies.

2.5 Scope of this report

This report aims to provide a synthesis of the work undertaken by the Paradigm project, the lessons learned along the way, and a series of recommendations arising from these. It is not an introduction to digital preservation, personal archives, preservation metadata or digital repositories, though some relevant concepts are introduced in it to facilitate elaboration of the project's activities. The report does not attempt to capture all the learning of the project. Archivists and technical staff faced with the preservation of digital archives can find more detailed and practical information in the project's Workbook, and in various project papers produced during the lifetime of the project.[12]

3 Findings

3.1 Archival principles and OAIS

The acquisition, processing and preservation of personal digital archives must continue to be underpinned by archival principles, so that creators and researchers have confidence in their continuing integrity. The OAIS model is highly compatible with archival principles, such as authenticity, context, original order and provenance. At the macro-level, the functions required to process personal digital archives are essentially the same as those required for traditional archival material (negotiating with depositors, accessioning, archival storage, administration, cataloguing and providing access, etc.); it is the micro-level activities that differentiate the management of digital archives.

In effect, traditional and digital materials of the same provenance may be accessioned together, but once they arrive at the research library they must be subject to different appraisal, metadata, archival storage and maintenance routines – digital archives should be directed to a secure digital preservation system designed for the needs of digital archival collections; traditional

[11] Cairo website http://cairo.paradigm.ac.uk

[12] The various papers produced during the Paradigm project are available at http://www.paradigm.ac.uk/projectdocs/papers/index.html

material should be subject to the usual assessments and BS 5454 archival storage.[13] At the point of cataloguing, the two media are reunited intellectually in a single finding aid for researcher access. The remainder of this report is arranged loosely around the lifecycle for personal digital archives developed by Paradigm over the course of the project, which borrows from traditional archival theory and practice and from new models, such as the OAIS reference model, that have emerged from digital curation and preservation research and practice.

3.2 Collection development policy and strategy

Most research libraries acquire personal archives from creators in the later years of their lives, or even after their death. Traditional collection development strategies have included:

- assessing archives that are offered as gifts, on loan or for purchase;
- actively monitoring auctions and sales;
- actively pursuing archival collections directly with creators who are well-established as historically significant.

The risk to digital archival materials is that degradation or obsolescence could render them inaccessible, or a complex processing prospect, long before they reach an archive via these routes. This could become increasingly true as those born today may use evolving technologies throughout their lives producing personal archives that represent several generations of computing technology.

The nature of the digital environment requires creators to becomes curators in their own right. This increases the likelihood that:

- creators enjoy short- to medium-term access to their own digital materials;
- research libraries accession full and complete personal digital archives that can be brought into the repository for long-term preservation and research use as personal archives.

In order to persuade creators that some measure of curation is required of them, an effort to engage with them earlier in the life cycle of their archive has been advocated. This is a period of learning for many creators of personal archives as much as it is for research libraries; it was easy to maintain access to their analogue letters, photos and documents, but many do not appreciate, or are only just beginning to learn, that more is required to ensure access to the digital equivalents of these materials; in addition, creators do not often benefit from the support of information and IT professionals to sustain their personal archives.

3.2.1 Strategies for hybrid personal archives

Paradigm considered a number of possible approaches to collection development in relation to archives containing digital materials. These range from familiar collecting approaches to those that are borrowed from the realms of organisational records management or eprint self-archiving:

- regular snapshot accessions – identifying and visiting creators at appropriate intervals and taking 'snapshots' of their records;
- post-custodial approach – identifying and supporting creators in the maintenance of archives that will be placed at the Library in future;
- traditional approach – waiting to approach, or be approached, by significant individuals later in life and monitoring sales;

[13] BS 5454:2000 *Recommendations for the Storage an Exhibition of Archival Documents*.

- transfer via retired media – identifying creators who will send devices and media that are no longer required to the Library;
- transfer via self-archiving interface – identifying creators who use a secure upload mechanism to transfer archival files to the Library as and when.

No approach is perfect. Pro-active engagement with creators earlier in the life cycle places new demands on research libraries in terms of selecting and maintaining relationships with individuals whose archives are now, or will be, historically valuable. They may also require even greater communication between research libraries about the specific foci of their respective collection development priorities. Approaches that involve waiting until digital archives or their creators are older risk loss of materials through physical degradation and obsolescence.

Proactive approaches are used successfully in other archival programmes. They involve selecting the archives of particular creators, because of the role they play and what their archive will document, rather than selecting archives based on an analysis of their content. This a kind of top down appraisal (variations explored in archival theory include documentation strategy, functional appraisal and macro appraisal). In politics, the application of these approaches is reasonably straightforward because the organisations and individuals are relatively small in number and well enumerated. It may be more complicated to apply this approach to fields such as science and literature, where identifying young scientists and writers whose archives have future research value may require a more retrospective viewpoint. Many advocates of macro-appraisal call for the involvement of expert researchers in collection development policy and strategy; in the Paradigm context this was provided by the project's academic advisory board of historians and political scientists. It was useful to discuss the policies and processes relating to collection development with researchers.

Recommendation 1: research libraries should involve researchers in formulating collection development policy and strategy and in identifying potential collections of interest.

Paradigm did not pilot each approach to collection development presented above. This was partly due to resource constraints and partly due to the preferences of participating politicians. More experience of applying these approaches to collection development over a longer period would be useful, including implementations in other domains, such as science and literature, which may give rise to different social, legal and technical issues.

Recommendation 2: more case studies that test different approaches to collection development are needed; from accessioning old media and devices to accessioning contemporary files.

Recommendation 3: case studies of working with personal archives and their creators in other domains, such as science and literature, are required in order to learn about the social, legal and technical issues pertinent to the whole range of creators of personal digital archives commonly collected by university special collections.

Neither did Paradigm explore the impact of digital archives on the trading of archives and manuscripts. Archival collections offered for sale could potentially accrue value if offered for sale later, when their creator is more well-established, but research libraries may need confirmation of the physical and moral integrity of a digital archive before purchase, and that they are buying the right to hold the sole research copy before investing resources in processing such material. This is particularly pertinent in the case of well-known writers, whose personal archives can sell for very large sums.

Recommendation 4: an analysis of the impact of the digital medium on the trade in archives and manuscripts would be of use to research libraries.

From the project's experiences of working with contemporary and older personal digital archives, Paradigm's conclusion is that a combination of approaches to collection development forms the optimum solution for a collection development programme. A variety of interventionist strategies should be used to support the varying needs of different individuals whose significance can be identified. Traditional approaches to selection must continue alongside interventionist approaches for those whose significance cannot be identified early-on in their career - significant personal archives that contain older digital materials cannot be rejected out of hand.

Among the implications of this conclusion are:

- resources are required to sustain several concurrent career-long relationships with creators;
- realistic and practical advice for creators must be developed and maintained, while avoiding liability;
- the expectations of creators must be managed appropriately;
- the skills and knowledge required to process a range of digital archives created in obsolete to cutting edge computing environments must be developed.

Paradigm's work on approaches to collection development, which has been published in the project’s Workbook, can be used by anyone wishing to consider the impact of digital archives on their collection development policy and strategy.[14]

3.3 Case study 1 - hybrid personal archives of working politicians

3.3.1 Approaching politicians

Working with contemporary politicians and their materials was the primary exemplar for Paradigm. The project aimed to work with individuals whose archives would fit with the existing holdings and collecting policies of the Bodleian and Rylands libraries. It was also felt to be important that the selected politicians represented a number of variables; Paradigm's academic advisory board were keen that the case studies should reflect a range of political roles and interests: ministers, MPs, peers and MEPs with local, national and international interests.

A number of politicians were invited to participate; responses to these invitations were not returned in all cases. In all, Paradigm worked with six contemporary politicians from the Conservative, Labour and Liberal Democrat parties; these included Members of Parliament, Peers and Members of the European Parliament, including politicians with Cabinet, Shadow Cabinet and important party political positions. This spread across political institutions and parties helped the exemplar to represent a diversity of working methods, organisational settings and the impact of individual attitudes towards IT, record-keeping and long-term archiving for historical purposes. It also provided a range of technical challenges.

We are very grateful to all the project participants for their cooperation.

[14] Paradigm Workbook http://www.paradigm.ac.uk/workbook/collection-development/index.html

3.3.2 Characterising the evolving nature of personal archives

To inform their discussion with participating politicians, the Paradigm project archivists first investigated the structure of politicians' personal archives more generally in an attempt to characterise hybrid personal archives. The project also considered the impact that current technologies might have on the content and structure of the average personal archive. This work was partly prompted by a concern that the continuing transition of digital mediums might obscure the archival qualities of new types of materials from creators, archivists and researchers alike, potentially meaning that important record series (e.g. blogs) are overlooked at the surveying and accessioning stages of the life cycle An awareness of the potential locations of personal digital archives is also useful, as they are highly likely to be scattered across multiple personal devices (PCs, laptops, phones and PDAs), network servers, online services and removable media.

The work initially involved examining the holdings of political personal archives at the Bodleian, Rylands and the Labour History Archives and Study Centre to compare traditional types of records, such as letters and journals, with their likely digital equivalents and seeking the input of Paradigm's academic advisory board. It was developed further in light of Paradigm's experiences of surveying the archives of participating politicians.[15]

Recommendation 5: digital archivists should maintain an awareness of the range of potential locations and mediums used for the storage of digital archives by their creators, and talk to creators about such materials to ensure that they are not overlooked.

3.3.3 The Paradigm records survey

Paradigm developed a records survey based on the findings of its work in order to characterise the contemporary personal archives of politicians; this evolved throughout the project as more experience of surveying and accessioning hybrid personal archives was gained. The records survey tool assisted the surveying archivists in identifying materials of historical value by assessing functions and roles, the nature of the records that document them, the vulnerability of records and their technical characteristics.[16] It also prompted creators to think about the historical value of their traditional and digital materials, as well as preservation-related issues.

Paradigm found that record surveying techniques are as useful for personal archives as they are for organisational archives - they allow archivists to record valuable contextual information about material that will be transferred both in the immediate and distant future, and therefore to prepare for its arrival. This is especially true in a digital context where more information is needed for initial transfer and processing. The Paradigm survey was sent to participating politicians in advance of a visit from the project archivists and completed by a mixture of interviews with creators, observation and records surveying. The information covered by the records survey is described in the following table:

Information	Purpose
Series of historical interest	**Identify location of material for accession; identify overlap between digital and paper records; descriptive overview assists in basic intellectual control of materials prior to cataloguing; informs initial appraisal decisions**
Quantity of material for accessioning	**Informs archivist's approach to accession**

[15] Paradigm Workbook http://www.paradigm.ac.uk/workbook/introduction/structure.html
[16] Paradigm Workbook http://www.paradigm.ac.uk/workbook/record-creators/surveying.html

Specification of hardware and software used to create records	**Informs archivist's approach to accession; recorded as preservation metadata**
Formats used	**Informs archivist's approach to accession; information about master and duplicate forms informs initial appraisal decisions; recorded as preservation metadata**
How/by whom records are created	**Recorded as useful contextual information for descriptive purposes, and provides important evidence of provenance**
Third party rights present in the material	**Inform administration of material; recorded as metadata**
Access restrictions required by creator	**Informs administration of material**
How records are managed and arranged by their creator	**Informs timing of accessions; informs subsequent archival arrangement of collections based on original order**

Paradigm supplemented the records survey tool with simple measures such as taking screenshots of file management tools and computer hardware details or generating textual files representing creators' arrangement of digital materials. Tips on how to do this are available in the Workbook.[17]

In total, these techniques were used to survey the records of working politicians at one charity office, two House of Lords offices, two House of Commons offices, three constituency offices and one European parliament office.

Recommendation 6: research libraries should use record surveying techniques when talking to creators and surveying materials; screenshots and directory trees can be very helpful. This information can be useful for a number of digital, and general, archive administration activities.

3.3.4 Selecting material from working politicians

Paradigm selected materials on the basis of perceived historical value. No effort was made to restrict on the basis of format as the project sought to embrace the widest range of personal archive data and format types. A total of seven accessions were made, comprising three accessions from one office, and a single accession from four others. Of these accessions, one was a hybrid accession, one a paper accession and five digital accessions. Some material accessioned represented a permanent addition to collections, other material was provided as a short-term deposit for the purposes of testing archival processes. Material included:

- Office documents – these made up a large proportion of each accession and included word-processed documents, spreadsheets, presentations;
- Images – these were mainly used for publicity materials, and often formed part of other documents or websites;
- Email – accessioned from Microsoft Outlook clients; this included a variety of associated attachment formats.

The types of materials accessioned varied from politician to politician; this was dependent on the records created in the first instance, and the willingness to place particular kinds of material in a Library. The project did not seek to acquire everything created digitally by an individual. While

[17] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html

some argue that this is possible in the digital domain, most archivists remain sceptical, wishing to eschew the burden of preserving excessive material indefinitely and of providing adequate researcher access to vast, low-value, collections. The project sought guidance from its academic advisory board on the characteristics of high quality personal archive collections for politicians. Important factors were:

- quality of the content – rich with good quantity and range;
- uniqueness of the material;
- focus on ideas;
- glimpses of high politics - material relating to the making of party policy on important public issues;
- glimpses of local activities;
- public materials that demonstrate how parties want to be viewed.

Additionally, the advisory board were asked about the kinds of digital personal archives that might be most valuable as research materials. Email and websites emerged as important series, though most types were considered to have some value. This potential value can also be dependent on the survival of related material that provides context. Researchers were less confident about the value of some digital materials, such as presentation slides and spreadsheets, that do not have obvious equivalents in traditional archives.

Recommendation 7: digital archivists should talk to researchers about personal digital archives to inform selection decisions and to increase awareness of the changing shape of personal archives.

3.3.5 Capturing and transferring personal digital archives

The means used to transfer personal digital archives from working politicians was informed by the information captured at the records survey stage, which ensures that the digital archivist arrives technically equipped to make the accession and knows what material is to be captured. In all cases transfers of digital archives were made by a visiting archivist, rather than remotely or by a third party. A transfer form was designed to record: the quantity of the archival material stored on each piece of storage media; its scope and content; supplementary information about rights and restrictions; and checksum information to verify the authenticity of the transfer process when the archivist returns to the library.[18]

Recommendation 8: research libraries should become familiar with means of proving authenticity in the digital environment, such as checksums and digital signatures.

In one case, working with a contemporary politician resulted in the transfer of older digital records stored on 3.5” floppy disks. In all other accessions, copies of digital materials that were stored on local or server hard disks were made. Accessioning copies allows the creator to retain the material while depositing a copy with the research library; this is a little different to analogue archival collections where it seems that 'unique' documents are transferred, though, in practice, these may be copies of various kinds too.

The media used in the transfer process varied according to the technical set-up of participants

[18] The transfer form is available as part of the Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html; information about checksums and digital signatures can also be found there.

and the volume of material being transferred; it comprised USB devices and CD-ROMs protected using biometric or password encryption.

Prior to transfer, it was necessary to extract some files from their storage locations. A particular case in point was email from a Microsoft Outlook client, which required the production of a how-to guide for obtaining files that could be transferred. Other kinds of archive that do not exist simply in obvious directory folder locations include PDAs and Smartphones, and any personal archive material stored with web services.

Recommendation 9: research libraries should become familiar with the extraction of digital archives from popular software, devices and web services.

Recommendation 10: research libraries should publish how-to guides for accessioning common kinds of personal digital archives for future reference and the benefit of others.

Later in the project, Paradigm was introduced to digital forensic tools, which have great potential for surveying and capturing personal archives on-site.[19] The project was unable to test the use of such tools in this context, but their design as means to capture evidence at crime scenes suggests that they meet the authenticity requirements for archives and have potential for surveying and capturing personal digital archives on-site, though some of the criminal terminology is potentially off-putting to archivist and creator.

Recommendation 11: research libraries should explore the potential of digital forensics hardware and software for survey and extraction of personal digital archives at the creators' premises and at the repository; a comparison of open source and proprietary solutions would be useful.

3.3.6 Lessons learned from working with contemporary creators

3.3.6.1 Access restrictions

Applying access restrictions of some kind for a period of years after deposit is common in personal digital archives; without such embargoes on access it is doubtful that materials could be obtained at all. It is difficult to make conclusive remarks about the impact that accessioning sooner after creation might have on access restrictions, save to say that material accessioned earlier in the life of its creator will be subject to legislative protections for longer while it is being administered by a library; its content might also be more sensitive at the point of accession, which may lead to management of archives that have more extensive closure periods. These costs may be offset by the benefits of accessioning materials in contemporary formats.

3.3.6.2 Impact of organisational change on personal archives

Most people are impacted by organisational change; politicians perhaps more than most. Some of these changes have an impact on personal archives and it is useful for archivists to be aware of predictable changes. In politics these include the general election (which limited Paradigm's contact with politicians for some months in 2005), staff changes and party reshuffles. Less dramatic events, such as the parliamentary recess, which gives staff time to appraise records,

[19] Thanks to Jeremy John of the British library for this.

can also impact on the historical archive. Other domains may be vulnerable to different changes, for example constant change in domains where project working is the norm.

Recommendation 12: research libraries should identify trigger points that may bring about the destruction of significant personal archives and aim to accession such materials if appropriate.

3.3.6.3 Third parties in personal archives

Personal archives documenting the professional lives of participating politicians are mostly created by aides, and include a great quantity of material created by, or in some way relating to, third parties. While true of traditional archives, the recentness of the material can dissuade creators from depositing some kinds of records. This has more to do with personal concerns than legal issues, as archives are permitted to hold such material under section 33 of the Data Protection Act (1998).

Political personal archives can be complicated by their potential overlap with records captured by public records legislation destined for long-term preservation at The National Archives. This requires research libraries to undertake additional measures in assessing whether the material should be open for research.

Recommendation 13: research libraries should aim to document third-party interests in archival collections they acquire so that these can be managed appropriately.

3.3.6.4 Time

Digital archivists met briefly with participating politicians, but dealt principally with staff responsible for creating and accumulating records associated with the professional aspect of a politician's life. While staff were receptive to archival and preservation concerns, their priorities are, understandably, day-to-day operations and the management of materials for short- and medium-term needs. Procedures designed for surveying and accessioning materials must therefore be as simple and efficient as possible, and guidance for creators must be realistic.

Recommendation 14: research libraries should use records surveying techniques to ensure that they are well-prepared to make accessions.

Recommendation 15: research libraries should be able to estimate the time needed to undertake an accession.

3.3.6.5 Storage pressures

Storage limits or declining storage capacity served as an impetus for large-scale record destruction; this was true of paper and digital archives.

3.3.6.6 A variety of collection development approaches are needed

All the project participants were different, and while sufficient similarities exist for the creation of generic approaches, research libraries must be flexible to cater for the variety of people, records and systems that are involved in working with multiple creators of personal archives.

Recommendation 16: research libraries should develop well-defined, but flexible, approaches to collection development and flexible accessioning procedures.

3.3.6.7 Successful development of personal archives relies on trusting relationships

Paradigm found that the reputations of the Bodleian and the Rylands were sufficient to allow the project to work with politicians, but evidence of the competence of the libraries to work with digital archives would have to be established over time. Curator-depositor relationships are very important and are often cultivated over many years before a collection, or the entirety of a collection, is deposited in an archive. This allows the building of trust and an appreciation of the archival tradition. The level of trust will impact on the materials placed at a library. It is unclear whether repository certification would currently assist in establishing credentials.

Recommendation 17: research libraries need to make a commitment to personal digital archives, moving this work from a project to a permanent footing. Contact with depositors whom research libraries wish to work with in the long-term should be made by permanent staff and form an integral part of their daily work.

Recommendation 18: research libraries should be able to reassure creators that material subject to embargo is held securely.

Recommendation 19: in order to encourage creators to deposit archives with research libraries sooner after their creation, research libraries should be able to reassure creators that private materials given to them are not subject to Freedom of Information provisions while the creator has placed an embargo on their access.

3.3.6.8 The potential of guidance for creators

Some personal archives, even copies, are unlikely to be placed in a research library until later in life. It is therefore important to increase awareness of archival and digital preservation generally, and specifically within domains and among individuals whose personal archives research libraries wish to collect. Individuals have access to a variable quality of technical expertise and IT support, which makes tailored advice preferable to generic advice in some cases. There is, however, a place for generic awareness of archival and preservation issues that are in the hands of creators and Paradigm has drafted some guidance of this nature.[20] This includes information on a number of relevant issues like backup, hardware maintenance, basic system administration, file formats and the selection of web services for managing personal archival material. A better basic knowledge of such issues will help individuals to manage their own digital assets; institutions such as the Digital Preservation Coalition may have a role here in getting the digital preservation message across to individuals via the media.

Recommendation 20: research libraries should offer generic or specific advice to the creators whose personal archives they are accessioning and/or aim to keep in touch with potential donors to monitor the technologies that will be coming to the repository in due course.

[20] Guidance for Creators of Personal Archives
http://www.paradigm.ac.uk/guidanceforcreators/index.html

Recommendation 21: more practical research into post-custodial collection development, where creators care for their own archives using guidance and systems provided by a research library, would be useful, though this may need to be undertaken over several years.

Recommendation 22: special interest groups (e.g. archivists working with literary, scientific or political personal archives) should raise awareness of digital preservation issues particular to specific domains by targeting their advice to those audiences.

Note: while advice can be useful, individual archivists and research libraries will need to ensure that this is provided with no warranty. Any developments to create a service for creators of personal archives should consider whether service level agreements and contracts are required and what resources are needed to sustain the service.

3.3.6.9 Agreements with depositors of hybrid archives

Paradigm produced a deposit agreement to cover the placement of exemplar materials for the duration of the pilot project, and a model agreement for the permanent placement of hybrid archives at a research library. These agreements help to establish the terms on which digital archives are placed with the research library, and ensure that both parties understand their obligations. Some aspects of the agreement are particular to hybrid and digital archives, such as securing rights for the library to undertake preservation actions on the material and securing the sole research copy of the archive. These agreements are available in the project's Workbook.[21]

3.4 Case study 2 - Web archiving pilot

The web archiving pilot, run alongside the UK Web Archiving Consortium (UKWAC) and London School of Economics Library, took snapshots of websites and blogs daily during the run-up to the General Election in 2005. Paradigm used HTTrack software, which produces a copy of the files that make-up a website in their existing structure and Adobe Acrobat Professional 7, which can be configured to create a copy of the website in many formats, including PDF.[22]

The project decided to investigate web archiving because personal blogs and websites are a significant personal record, and Paradigm wanted to assess whether collecting such materials as part of a personal archive could be incorporated into the activities of research libraries. Gathering reasonable quality snapshots of these websites was fairly straightforward, but like other web archiving initiatives Paradigm found that obtaining permission to archive websites was a time consuming, and often fruitless, task. Obtaining permission from the politicians with whom Paradigm had existing relationships was more successful than approaching politicians with whom there had been no previous contact, although permission was obtained from some additional politicians. This suggests that archiving the website of an individual as a part of their personal archive is a good strategy.

Preserving and providing access to collections of personal website snapshots is more complicated than acquiring them, as websites are complex digital objects consisting of several inter-related digital files in numerous formats. Tools from the web archiving community may make their administration easier in future,[23] but even so, research libraries may prefer to

[21] Terms of Agreement for Personal Archives http://www.paradigm.ac.uk/workbook/accessioning/documentation/index.html

[22] HTTrack http://www.httrack.com ; Adobe Acrobat Professional (now at v. 8) http://www.adobe.com/products/acrobatpro

[23] The International Internet Preservation Consortium is developing a Toolkit to support webarchiving http://netpreserve.org

cooperate with a central web archive to ensure that snapshots of a particular individual's website are included in a national archive, rather than undertake this activity themselves. UKWAC could fulfil this role, but how this consortium will operate in the future, and whether, and at what cost, it might undertake to archive websites at a frequency specified by external organisations is uncertain.

Recommendation 23: research libraries should keep a watching brief on developments in webarchiving initiatives and technologies.

3.5 Case study 3 - Digital archaeology pilot

The archive of former Cabinet Minister Barbara Castle - a more traditionally deposited collection - is placed at the Bodleian Library. This includes some 500 boxes of traditional archive materials, two personal computers and 3" and 3.5" disks. Paradigm decided to work with this older personal digital archival material for the experience of working with personal archives created in less contemporary formats and because there was no other obvious means of processing it. Older computers and Amstrad disks are likely to be included in other personal archives that will come to research libraries, so developing and documenting the means to process them seemed a useful endeavour. It is not always possible to send digital materials to third-party data extraction services due to privacy concerns; doing so may require permission from the donor.

Recommendation 24: if research libraries intend to use third parties to process obsolete digital archives they should include permission to do so in deposit agreements.

Because this mode of working means that the Library has the sole copy on older, potentially fragile, media, the first task is to authentically extract the data and ensure that backup copies are made before working at accessing and appraising the content. The hard disks of the PCs were extracted from their PC housings and Paradigm worked with Jeremy John, Digital Curator at the British Library, who demonstrated how digital forensic hardware and software can be used to extract an image of a hard disk.[24]

Images of hard disks are interesting in that they contain the software used by the creator as well as files they have created. It is unclear whether research libraries can retain and use proprietary software found on such disks for archival purposes; while commercial interest may seem to have expired, the restrictive licences applied typically have no end-date and most do not permit the transfer of the licence from its original licensee.

Recommendation 25: the preservation community should lobby for changes that would see vendors rescind commercial interests in obsolete software, or change the law so that software licences contain an expiry date after which their use is not constrained by the terms of the original licence.

Digital forensic software uses known file filters to identify non-archival files, such as software (based on checksums of such software maintained at the National Software Reference Library,[25]

[24] Disk images can be taken in the open, documented format (dd), and in tool-specific proprietary formats. An overview of digital forensic tools for this purpose is available at the *Forensics Wiki* http://wwwforensicswiki.org/wiki/Tools:Disk_Imaging

[25] NSRL is maintained by NIST http://www.nsrl.nist.gov/. A third-party search interface to the database is available: http://ionrift.ath.cx/nsrl/.

etc.). It is useful to be aware of the typical directory locations for storing files used by certain operating systems and application software, although the surveying capabilities of forensic tools does help with this.

Paradigm found that forensic hardware and software is best suited to more modern computing environments; for older materials it is often a case of recreating an environment in which the data can be extracted. For the 3” Amstrad disks Paradigm undertook some research, greatly facilitated by Amstrad enthusiasts and their websites, and assembled a toolkit from eBay for extracting the data; the toolkit approximates to that used by Amstrad PCW users wishing to migrate their data to an IBM PC platform in the 1980s.

Recommendation 26: websites that record information about extracting data from past and present personal computing environments should be submitted to UKWAC or the Internet Archive for archiving by research libraries.

Recommendation 27: research libraries should create how-tos for the extraction and migration of common obsolete digital media and file formats.

Recommendation 28: research libraries should be able to access preserved and working examples of the hardware and software needed to extract data from older computing environments.

Recommendation 29: research libraries should promote the use of open standards that pose fewer problems for the preservation community and others.

3.6 Processing digital accessions

3.6.1 Initial processing

The processing of newly-arrived born-digital archives involves a number of steps before the material can be submitted to a digital repository for archival storage. The conclusion of this stage is an Archival Information Package,[26] that contains all the preservation metadata needed to support the repository in preserving the digital archive material. This is where the real departure from traditional archival practice begins. A series of activities commence, the exact details of which are dependent on the nature of the materials.

Activities that embrace the hybrid archive

Some activities are common to the hybrid archive as a whole:

- gaining intellectual control over the material;
- filing paperwork – agreements, correspondence;
- adding an entry to the library's accessions register that captures high level administrative metadata.

[26] The Archival Information Package is a core part of the OAIS' information model, see section 2.2 of the OAIS model (ibid).

Activities particular to the digital materials in the archive

When the digital component of the archive is transferred to a digital preservation service, the following digital-specific activities should be performed:

- creation of backup and working copies (using off-site and fireproof storage);
- health check for viruses and to determine digital object integrity and validity;
- technical appraisal of the material and how it must be processed;
- format transformation of formats not supported by the repository;
- creation of preservation metadata, including allocation of low level persistent identifiers;
- submission of Archival Information Packages to the digital preservation repository.

As part of the Paradigm prototype workflow and repository architecture, a standalone accessions area was created where incoming digital materials could be subject to this initial processing. Paradigm needed to ensure that the creation of copies, whether deliberate or incidental, was strictly controlled. For simplicity and security, the project created an accessioning and archiving environment that was isolated from the network. This removed concerns about the security of data in transit over the network, attacks on the repository over the network, accumulation of sensitive materials on staff workstations and exposure of material to the sneakernet.[27] To ensure the security of material, in accordance with Data Protection principles, we physically limited access to project staff. In a production environment, and as the quantity of digital archive materials grows, additional staff may need access to closed material in the preservation repository to undertake enquiry, freedom of information related, planning or cataloguing work, so some kind of secure access over a network will be required in future.

Recommendation 30: research libraries should investigate how best to provide secure access to archival storage for archival staff over a network.

Recommendation 31: research libraries should be aware of the incidental copies of digital materials made by computers and how digital material is securely deleted.

In the accessions area material was checked for viruses and subjected to analysis by a range of digital curation tools, and structured XML metadata was created to form part of its Archival Information Package.

Some of these activities require the development of policy and procedures relating to the treatment of formats. Preservation repositories need documented business rules for how they will treat each format, e.g. whether it is migrated and what metadata must be captured about it. This enables materials to be treated in a consistent fashion allowing the repository to reap economies of scale in its future operations and facilitating the training of new personnel. The development of such business rules was beyond the scope of the Paradigm project.

Recommendation 32: research libraries must develop business rules for the treatment of data types and data formats so that digital preservation can be as consistent, predictable and economical as possible.

Recommendation 33: a central repository providing information, and perhaps access to tools, needed to identify good quality migration

[27] 'Sneakernet' is a term that describes the use of removable storage devices for moving digital materials between systems.

pathways would be helpful in devising business rules for the treatment of formats submitted to research libraries.

3.6.2 Computing environments and file formats encountered during Paradigm

One of the key pieces of metadata needed for preservation is information about the computing environments and file formats of digital materials. A range of tools are available for:

- identifying the formats of files;
- validating that files conform to their format specifications;
- extracting or generating basic preservation metadata.

The registries and tools tested by the Paradigm project included:

- DROID tool for identifying formats;[28]
- PRONOM registry for access to information about formats;[29]
- FILEEXT registry for basic format information;[30]
- JHOVE for validating formats and generating technical metadata;[31]
- National Library of New Zealand Metadata Extraction Tool;[32]
- Wikipedia for format related information.[33]

Formats accessioned from working politicians, that encoded files created in the last five years, were typical of contemporary office environments. All were accessible using contemporary technologies, though some of these are proprietary and restrictive. The relative ease with which this material can be surveyed and accessed at the time of accession demonstrates the benefits of acquiring archival materials sooner after creation.

3.6.2.1 Computing environments of contemporary politicians

Common formats in the accessions from contemporary politicians included:

- several versions of Adobe PDF;
- several versions of MS Word;
- several versions of MS Excel;
- several versions of MS Powerpoint;
- ASCII plain text;
- personal store (.pst) files extracted from MS Outlook;
- various image formats, including versions of jpeg, gif and tiff;
- little in the way of audio and movie files.

These derived from a range of 32 bit operating systems:

[28] DROID http://droid.sourceforge.net
[29] PRONOM http://nationalarchives.gov.uk/pronom
[30] FILEEXT http://fileext.com
[31] JHOVE http://hul.hardvard.edu/jhove
[32] NLNZ Metadata Extract Tool, currently being redeveloped for release under an open source licence. See http://www.natlib.govt.nz/about-us/current-initiatives/metadata-extraction-tool
[33] Wikipedia http://www.wikipedia.org

- Microsoft Windows 95;
- Microsoft Windows 98;
- Microsoft Windows 2000;
- Microsoft Windows XP.

Storage media encountered included:

- 3.5” floppy disks;
- USB keys;
- CD-Rs;
- DVD-Rs;
- Hard disks;
- Networked storage.

3.6.2.2 File formats found during the webarchiving pilot

The webarchiving pilot was small scale and did not include any particularly complex web material. The websites were typical of those established by politicians at the time of the pilot. The material in the website harvests included:

- various versions of html;
- Cascading Style Sheet (CSS) files;
- various image formats, including versions of jpeg, gif and Windows bitmap;
- various versions of Adobe PDF;
- Javascript files;
- Macromedia Flash;
- MS Word for Windows 97-2003.

3.6.2.3 Computing environments represented in the Barbara Castle archive

The material in the Barbara Castle archive included:

- Locomotive Locoscript files;
- Wordperfect 5.1 for DOS files;
- Microsoft Word for Windows 6.0/95
- MS Excel for Windows 7.0/95

and was created on 8 bit, 16 bit and 32 bit operating systems:

- CP/M;
- Microsoft Windows 3.1;
- Microsoft Windows 95.

It arrived at the Library on:

- 3” Amsoft disks;
- 3.5” floppy disks;
- two IDE hard disks.

Paradigm found that the registries and tools available from the digital curation and preservation community had focused on popular contemporary proprietary formats or on formats used in digital library collections. Formats used in personal computing environments that are older or more obscure are not so well supported. This could be resolved over time by the contribution of relevant information from research libraries processing materials in these formats.

Recommendation 34: research libraries should contribute information to file format registries so that their content can be developed to include a wider range of the formats typically found in personal digital archives.

Recommendation 35: the digital preservation community should raise awareness of file format registries with open source and proprietary software developers.

3.6.3 Appraisal

In the case of the exemplar archives accessioned from working politicians, a reasonable amount of appraisal had taken place long before the material arrived at the library. Individuals whose archives were considered to be historically significant had been identified, and record series of value had been selected on-site and transferred to the library. Appraisal at the macro level was therefore largely complete. At the library, what remained to be done was some measure of bottom-up appraisal, which focuses attention at the item level. This generally comprises the identification and elimination of duplicate files; thanks to checksumming techniques, which can pinpoint exact copies, this is simpler in the digital environment than in the analogue.

The appraisal of digital materials that arrive at an archive via more traditional routes present different challenges. Before the content can be appraised, the files must be liberated from the media on which they reside and there is a degree of uncertainty as to whether the process will be worthwhile, as the likely contents of media cannot be ascertained and its physical integrity is not necessarily obvious on visual inspection.

Sometimes data extraction involves recreating whole digital environments, as in the case of Barbara Castle's 3” Amsoft disks. Recreating older technical environments is initially time consuming, as it involves some learning and the acquisition of necessary components. If, however, research libraries continue to receive digital media created in 1980s computing environments over the next few years, surely this can become a familiar aspect of the digital archivist's work: the techniques, equipment and procedures developed to support it can be documented and re-used in such a way that the up front investment in developing such capacity is returned as successive personal archives are processed.

At other times data extraction is simpler. Computer hard disks with recent operating systems do not pose great problems, though decisions about what to do with the hardware and the software will be needed, as will retention decisions about the various storage media on which personal archives are deposited once their contents are extracted and placed in the managed environment of the digital preservation repository.

Further appraisal based on the content and value of the material may optionally be undertaken as part of the cataloguing process when the archive is being prepared for researcher access. This requires the cataloguing archivist to understand the content of the material, sometimes by relating it to other material in the archive. Paradigm evaluated appraisal guidance for traditional archival collections, and explored the use of tools and techniques for appraisal, producing some tips on how to do this in its Workbook.[34] Unsurprisingly, the project archivists found it much easier to appraise structured archives than disorganised ones. The trend towards using search

[34] Paradigm Workbook http://www.paradigm.ac.uk/workbook/appraisal/index.html

technologies in place of file structures may frustrate the attempts of archivists to arrange and describe larger digital personal archives, although there is also a trend toward the annotating and labelling of materials with keywords that could facilitate appraisal if this metadata can be preserved alongside the material it describes. The use of labels and tags in services such as Gmail, You Tube and Flickr, and in desktop software, are all means by which creators can enrich their data, but extracting the data for placement at an archive with this metadata incorporated is not generally possible.

Recommendation 36: research libraries should monitor trends of functionality and use in the arrangement and use of personal digital materials.

Recommendation 37: research libraries should encourage creators to organise materials rather than rely on searching technologies.

Alternative approaches to the arrangement and description of archival materials may be to derive indexes from keywords and personal names found in textual materials, though this does not provide means of access to digital image, sound or movie materials, and does not provide the contextual overview of a traditional archival description.

3.6.4 Disposal

The deposit agreement created for the Paradigm project included a clause that allowed the libraries to destroy or return material that was not archival. In practice, it is more likely that archivists will destroy materials that were accessioned as copies of the creator's holdings. Disposal policies and procedures, based on an understanding of secure deletion for digital materials, will be required.

3.6.5 Preparing preservation metadata in Archival Information Packages

Paradigm piloted the creation of Archival Information Packages that contained preservation metadata necessary for the preservation of its exemplar digital archives for submission to its prototype Fedora-based digital preservation repository. The METS[35] metadata standard was used by the project for this purpose. Benefits of METS include:

- it is well-supported in the digital library community and by digital repository software, such as DSpace and Fedora;
- it allows the combination of several kinds of metadata encoded in different XML standards within a single XML file, thus allowing all the metadata about an object to be packaged together;
- it is flexible enough to deal with a wide variety of materials; this is particularly useful in the context of personal archives which typically contain a wide range of object types;
- it allows for the indefinite extension of the metadata held in the METS file – additional, or new versions of, metadata can therefore be added to the METS file over the life of the material it documents without overwriting previous metadata;
- it provides mechanisms for recording the structure of digital materials, which can be used to record the original order of archival accessions;
- its use in preservation is being explored by many digital curators.

[35] METS http://loc.gov/standards/mets

Disadvantages include:

- METS is highly flexible and the variety of implementations arising from this affects interoperability;
- existing METS profiles (the publication of profiles is an attempt to limit the variety of implementations and to provide examples) pertain, by and large, to access-related rather than preservation-related projects.

For preservation metadata applicable to all kinds of digital object, such as size and file format, Paradigm worked with PREMIS,[36] which was released during the lifetime of the project. Some metadata extraction tools produce the metadata defined in the PREMIS Data Dictionary, but no tool encodes this in PREMIS' XML schemas at present. While PREMIS deliberately eschews mandating how PREMIS-compliant metadata should be stored, it would facilitate staff understanding and system interoperability if metadata tools could export their results in PREMIS XML rather than native schemas.

Recommendation 38: research libraries should aim to record PREMIS metadata using the PREMIS XML schemas within METS files. This should increase the understandability of the metadata and facilitate interoperability.

Paradigm also considered the various ways of encoding rights metadata, including METSRights[37] and the Rights entity in PREMIS. It is hoped that the Gowers[38] recommendation to allow format shifting of archival materials will be enacted in the near future, as recording of rightsholder details (such as provided by METSRights) at the item level is impossible. PREMIS Rights only allows only for the recording of permissions granted by rightsholders; this does not cover the rights environment in which archivists work where the majority of rights are held by, sometimes unidentifiable, third parties who have given no permission to undertake preservation actions, and therefore such actions must be undertaken within the confines of restrictive legislative provisions.

For preservation metadata that is specific to the technical attributes of digital object types (e.g. still image, text, audio and video) the project briefly explored a number of metadata standards, including MIX[39] for images, textMD,[40] audioMD[41] and videoMD.[42] Most of these are either developed by or maintained at the Library of Congress. This is one of the more complex areas of digital preservation, as few tools are available to create this technical metadata, and comprehensible introductions to the metadata, how it can be derived and why it should be recorded are not available for most types of object.

Recommendation 39: more accessible background information for technical metadata standards is needed by research libraries.

Paradigm's work on learning about and testing the potential of METS, PREMIS and technical and

[36] PREMIS http://loc.gov/standards/premis

[37] METSRights http://www.loc.gov/standards/rights/METSRights.xsd

[38] See Recommendations 10a and 10b of the *Gowers Review of Intellectual Property* (December 2006); http://www.hm-treasury.gov.uk/independent_reviews/gowers_review_intellectual_property/gowersreview_index.cfm

[39] MIX http://www.loc.gov/standards/mix

[40] textMD http://www.dlib.nyu.edu/METS/textmd.xsd

[41] audioMD http://www.loc.gov/rr/mopic/avprot/DD_AMD.html

[42] videoMD http://www.loc.gov/rr/mopic/avprot/DD_VMD.html

rights metadata standards for personal digital archives is documented in the Workbook.[43]

METS metadata packages were created by assembling the various XML outputs of tools, and hand-crafting some metadata based on human or tool assessments of the sample digital archives, which were combined in METS files using XML editors or via the Fedora client software. This was immensely difficult and laborious. While digital archivists should understand the metadata that is needed to preserve digital objects, an interface to coordinate the activities of multiple metadata extraction tools, which can relieve the burden of compiling METS files, is needed for the activity of crafting preservation metadata if this is to scale. Creating preservation metadata for personal digital archives collections is especially complex because of:

- the relationships between component parts of the archive which are so important;
- the wide range of digital object types present, some simple (e.g. gif image) and some complex (e.g. personal mail store extracted from MS Outlook), requiring different technical metadata.

Recommendation 40: research libraries need a single user friendly interface that combines the functionality of a number of metadata extraction tools to create METS-based Archival Information Packages.

Recognising that a practical solution to a problem of this scale was beyond the resources of Paradigm, the project partners created a new software development project in partnership with the Wellcome Library. The JISC-supported Cairo project is to develop a tool for creating Archival Information Packages for personal digital archives.[44]

3.6.6 Persistent identifiers for personal digital archives

Paradigm also investigated the implementation of persistent identifiers (often known as PIDs) necessary in the context of personal digital archives and the schemes available for persistent identification. Preservation of paper materials prior to cataloguing is undertaken by high level activities, such as configuring and monitoring the storage environment, so the allocation of granular identifiers can normally wait until the cataloguing stage. Preservation of digital materials takes place at a lower level than with traditional archives, and preservation activity, which must be documented in the item's audit metadata, may be required prior to the cataloguing stage. Digital objects therefore need a persistent identifier from the moment they are ingested into a digital preservation repository. When objects are migrated to newer formats these new instances also require identifiers which can be used in metadata to relate them to the object from which they are derived for various other preservation and access needs. Persistent identifiers are therefore an important part of any digital preservation system. Paradigm's investigation found that the technical system adopted for PIDs mattered less than a knowledge of what needed to be identified and why, and a commitment to the ongoing maintenance of the system chosen. A Workbook section is available which introduces the project's learning about PIDs and the ways in which they might be used in preserving and providing access to personal digital archives.[45]

Recommendation 41: research libraries should develop a model for persistently identifying the component of personal digital archives carefully and commit to the maintenance of the system chosen.

[43] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html
[44] Cairo project http://cairo.paradigm.ac.uk
[45] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html

3.7 Digital preservation repository

Paradigm created another standalone system for its preservation repository. This was a 'dark archive' designed purely for archival storage that should be subject to preservation monitoring and actions, and secured to protect material contained in it that may be sensitive and subject to embargo. It was envisaged that once archives were to be opened for public access, 'access copies' of digital archives would be supplied to an accessible repository, while preservation of the archives would continue in the preservation repository. This separation of functions allows the creation of a secure preservation repository concentrated on preservation needs, and an access repository optimised for access functions. Paradigm found it difficult to address such issues as the likely primary and backup storage required over the coming years; the quantity of incoming traditional archives can be variable in collecting institutions, but there is, at least, some historical data available for predicting needs.

Recommendation 42: research libraries should monitor trends in the growth of personal digital archives to plan for storage needs. Growth includes incoming collections and their metadata, increasing metadata for existing archives and extra files created by format shifting, etc.

The prototype digital preservation repository received material processed in the accessions area, and therefore included the preservation metadata required at the point of ingest into the system. Paradigm tested DSpace and Fedora as potential candidates for repository system software, and opted for Fedora. A comparison of the two can be found in the Paradigm Workbook.[46] In brief, we chose the Fedora system because:

- it was proven to scale well;
- the system was being used by other users for complex collections;
- Fedora is agnostic in its approach to file formats; this suited the context of personal digital archives that contain a wide variety of digital object types;
- the system has an open architecture – it is built around the idea of a service framework, allowing users to contribute to its functionality without impacting the core repository's code;
- the system had good audit capabilities;
- the system was introducing repository and lower level security controls (now implemented);
- developments related to preservation were very much on the development roadmap.

At present no repository system can offer an end-to-end preservation solution and the potential of a system like Fedora in undertaking some of the relevant activities will be very much dependent on the implementation of the system. There are signs that in the not too distant future mechanisms for obsolescence monitoring, including interfacing with file format registries, will be available.[47] Fedora has also implemented preservation-friendly features in its latest release (2.2), such as recording checksums and performing integrity monitoring of digital objects and metadata.[48]

One of the greatest difficulties with Fedora is also one of its strengths – the very flexibility it has means that the system deployed “out of the box” lacks constraints and its user interfaces are very limited. This is because the developers expect implementers to develop what suits their, sometimes very different, needs. Uptake of the system would be greatly facilitated by the

[46] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html

[47] Wiki page of the Fedora Preservation Services Working Group http://www.fedora.info/wiki/index.php/Working_Group:_Preservation

[48] Fedora 2.2 Release Notes http://www.fedora.info/download/2.2/userdocs/distribution/release-notes.html

provision of domain-specific Fedora configurations, data models and interfaces. Such interfaces as VITAL, VALET, Fez and Elated have been created for eprint repositories,[49] but these cannot handle the variety of materials, the relationships, and the preservation metadata required for a digital preservation repository for personal digital archives.

Recommendation 43: research libraries should develop data models, and interfaces based on those models, for using Fedora as a preservation repository for personal digital archives. Fedora for archivists out-of-the-box would make its use a possibility at institutions with less technical support which might otherwise have to develop divergent local custom solutions.

As part of the work on repositories, the Paradigm Workbook contains installation instructions for DSpace and Fedora, tests of Fedora's DirIngest service and an authentication and authorisation model for user access and permissions implemented in Fedora using the XACML policy language.[50]

3.8 Preservation strategy, monitoring and actions

Paradigm's findings are that preservation monitoring will be a combination of local and global technology watch. Repositories should contain metadata that allows them to characterise the technical nature of their holdings and to plan preservation actions, and the development of preservation capabilities, accordingly. Some local choices may be dependent on development trends in the preservation community as a whole; arguably, single institutions cannot support preservation strategies that are divergent from community trends.

Recommendation 44: research libraries should perform local technology watch based on a knowledge of their holdings, and remain aware of technology watch services available in the digital preservation community and of digital curation and preservation trends.

The Paradigm Workbook contains an introduction to preservation strategies considered from the context of those responsible for the preservation of personal digital archives.[51]

3.9 Access to personal digital archives

3.9.1 Descriptive cataloguing

Paradigm used its exemplar archival collections to examine the descriptive challenges posed by hybrid archival collections, and how Encoded Archival Description[52] (used by the Bodleian and the Rylands to encode ISAD(G)2[53] compliant archival descriptions for existing archives) might be used for cataloguing hybrid archival collections, and how METS might be used to support their display and navigation in a repository. This work raised a number of issues, such as:

- the need for expressing the balance between traditional and digital parts of a personal archive;

[49] The Fedora Tools webpage provides links to these, and other, tools developed by the Fedora community http://www.fedora.info/tools

[50] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html

[51] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html

[52] EAD http://www.loc.gov/ead

[53] ISADG(2) – International Standard for General Archival Description, Second edition http://www.ica.org/biblio.php?pdocid=1

- the need for devising appropriate measures for meaningfully describing the extent of digital materials;
- the potential for linking from EAD to digital object and the level of cataloguing at which this should be done;
- the kind of metadata researchers would need to establish the authenticity of migrated objects.

The project also investigated the potential of the Archives Hub[54] as a means of distributed access to personal digital archives, by talking to the Archives Hub team and asking them to complete a questionnaire. The findings of this work were:

- the Hub is planning to develop support for linking to digitised archives using METS; this could form a useful basis for the technical work involved in linking from Hub EAD descriptions to born-digital archives;
- the Hub does not envisage hosting born-digital archives, or assuming responsibility for their preservation. It would purely link from EAD to the representation of the born-digital object in the research library's own repository;
- data protection and copyright constraints mean that whole collections of born-digital personal archives are unlikely to be mounted in freely accessible online repositories for a number of years and therefore access via the Hub would require some kind of central authentication and authorisation regime; the complexity of access conditions present in archival collections means that they are unlikely to be compatible with such a regime.
- the Hub team play a role in informing the archival profession of issues relating to professional development, and could become a centre of advice on digital preservation issues for archivists.

The Paradigm Workbook contains an introduction to relevant standards for archival description and an overview of the impact of the hybrid environment on archival arrangement and description practices based on the exemplar collections accessioned by the project.[55]

3.9.2 Access requirements for researchers

The period of transition for the researchers who use personal archives has yet to begin in earnest, as most of the personal digital archives accessioned by research libraries to date are yet to be opened for research use. In future, researchers will need to be familiar with the cultural and material nature of digital archives such as email, personal blogs, digital photo and video collections, office documents and digital diaries, just as they currently are of ancient papyri, medieval manuscripts, Victorian letters, albumen prints, desk diaries and travel journals.

There will be different techniques for judging authenticity, dating materials and identifying authors and different access techniques made possible by digital archives. These access possibilities, such as free text searching and sorting, will enable researchers to interrogate personal archives in ways that are impossible for traditional archival materials, but researchers will need to develop an awareness of these possibilities; the possibilities of analysing the rhythms and relationships found in email archives is particularly interesting.[56] Researchers will also need a basic understanding of digital preservation, so that they can interpret the digital provenance of digital archives that have been subject to preservation actions, such as file format migrations.

[54] Archives Hub http://www.archiveshub.ac.uk

[55] Paradigm Workbook http://www.paradigm.ac.uk/workbook/index.html

[56] Adam Perer, Ben Shneiderman and Douglas W. Oard 'Using Rhythms of Relationships to Understand Email Archives' http://hcil.cs.umd.edu/trs/2005-08/2005-08.html

Recommendation 45: research libraries should engage with the research community to promote a wider understanding of digital preservation issues and their impact on the primary sources that will be available in future years.

Recommendation 46: research libraries should investigate how access to personal digital archives will be delivered.

3.10 Legal issues

An analysis of the impact of key legislation (Freedom of Information, Data Protection Act, Public Records Act, intellectual property rights. etc.) on collecting, preserving and providing access to digital archives was undertaken as part of the project. This area is documented in the Paradigm Workbook.[57] One of the key concerns in a digital environment is understanding the impact of legislation on how materials that are still subject to privacy and intellectual property rights legislation can be delivered to users for reading, and what regimes for supplying copies will be acceptable. Paradigm's work has raised some of these issues, but the input of legal advice would be helpful in determining how libraries should proceed.

Recommendation 47: research libraries should seek legal advice about the provision of digital materials both in a stand-alone reading room environment and over networks, and about the supply of copies for research use.

3.11 Skills needed to work with digital archives

The Paradigm project has demonstrated that the skills needed to work with personal digital archives are diverse. They can be distributed across a team, across departments and even across institutions. Among them are:

- an understanding of archival principles;
- an ability to analyse business functions and activities and the impact of technology on how these are undertaken and what records are created;
- an understanding of legislation applicable to the information environment;
- an ability to engage with emerging and older tools and technologies that may be unfamiliar or poorly documented;
- an understanding of Information Technology – hardware, software, media, formats, etc.;
- an ability to identify evolving social and technical issues in the changing IT landscape;
- an understanding of XML technologies;
- system administration skills;
- an ability to develop policy;
- an ability to model procedure and workflow;
- an ability to analyse formats, tools, metadata;
- an understanding of researcher needs;

[57] Paradigm Workbook http://www.paradigm.ac.uk/workbook/legal-issues/index.html

- management skills;
- an ability and willingness to continue developing knowledge and skills on an ongoing basis.

The following recommendations are made with regard to bridging the skills gap apparent in the archival community at present. The archival profession needs specialists in the management of digital materials, just as it needs specialists in other areas. That is not to say that every repository needs a digital archivist, any more than every repository needs a medieval palaeography expert, but those that expect to collect important digital archives on any kind of scale should expect to engage with a specialist, even if that specialist exists outside their organisation. While digital curation practices mature they are likely to be rough around the edges and unfriendly to newcomers.

Those that do become digital curation specialists need support from the larger community; collaborations with similar institutions can help here (as it has for the Paradigm participants) as can the invaluable training, resources and support offered via the Digital Preservation Coalition (DPC) and the Digital Curation Centre (DCC).[58] Naturally, the availability of specialists is no panacea; digital archives are a distributed responsibility, and a better general level of awareness throughout the archival profession, and beyond, is also needed. The archival community can build this capacity in a number of ways.

Recommendation 48: research libraries should aim to progressively build experience in working with digital archives through direct engagement with creators and their personal archives.

Recommendation 49: research libraries should take advantage of the offerings of organisations such as the DPC and DCC.

Recommendation 50: graduate traineeships for new entrants into the archival profession, and archival courses and modules that allow students to specialise in born digital archives should be created.

4 Acknowledgements

The Paradigm project wishes to extend its thanks to all who assisted this project in achieving its aims.

4.1 Project team

Project Manager and Digital Archivist: Susan Thomas

Digital Archivist: Janette Martin

Developer: Renhart Gittens

Digital Archivist: Fran Baker

4.2 Funders

Thanks go to the JISC and Oxford's Research Development Fund (now the John Fell OUP Research Fund) for funding the work of this project, and in particular to Helen Hockx-Yu, Programme Manager at the JISC, for her support.

[58] Digital Preservation Coalition ; Digital Curation Centre.

4.3 Steering Group

Stella Butler: Head of Special Collections & Principal Keeper, John Rylands University Library

Chris Fletcher: Head of Western Manuscripts, Bodleian Library

John Hodgson: Keeper of Manuscripts & Archives, John Rylands University Library

Helen Langley: Head of Modern Political Papers, Bodleian Library

Richard Ovenden (Chair): Head of Special Collections & Associate Director

Michael Popham: Head of Oxford Digital Library

Emily Tarrant: Conservative Party Archivist

4.4 Academic Advisory Board

Lawrence Goldman (Chair): Editor in Chief, Oxford Dictionary of National Biography

Simon Bailey: Keeper of Oxford University Archives

Dr Martin Conway: Lecturer in Modern History, University of Oxford

Dr John Davis: Lecturer in Modern History, University of Oxford

Professor Steven Fielding: Professor of Contemporary Political History, University of Salford

Dr Alex May : Research Editor, Oxford Dictionary of National Biography

Dr Kevin Morgan: Department of Government, International Politics and Philosophy, University of Manchester

4.5 Participating politicians and their staff

Thanks go to the politicians and their staff, who made this project possible by willingly giving us their time and cooperation.

4.6 Others

Thanks to the nameless many, within and beyond Oxford and Manchester, that contributed to this project, by providing advice, comment and encouragement.