DIGITAL PRESERVATION

an introduction to the standards issues surrounding the deposit of non-print publications

A study conducted by
Book Industry Communication
on behalf of the
British National Bibliography Research Fund

Mark Bide, Liz Potter, Anthony Watkinson

September 1999

BIC

BOOK INDUSTRY
COMMUNICATION

Contents

1   Executive Summary

1. This study is intended to provide a focus for the discussion of the technical and procedural
    challenges involved in developing a national archive of electronic publications in the UK. It
    is intended that it will assist in informing the code of practice for the voluntary deposit of
    certain types of electronic materials, which is currently being discussed by representatives of
    the Publishers Association and of the UK copyright libraries.

2. The aim of the study is to provide a useful starting point for those readers who are looking for
    a way in to the extensive literature on the topic. The study itself differs from much of that
    literature in a number of ways. It focuses on digital preservation in the context of national
    deposit, rather than on the impact of digital preservation on the full spectrum of library
    services. It is essentially concerned with practical implementation within a framework already
    existing, rather than with the development of new structures. In addition, it attempts a cross-
    sectoral approach, examining the needs and concerns of both publishers and libraries.

3. We begin by establishing the need for digital preservation. Our ability to continue to use
    digital information is threatened by the limited life of the media on which it is distributed and
    by the rapid obsolescence of the particular software and hardware required to access and
    interpret it. The development of a national published archive of electronic material is one way
    in which we can address our need to preserve the growing amounts of culturally significant
    information which is available only in electronic form.

4. We consider the definition of electronic publications, which is central to the technical
    challenges associated with preserving them. We outline projects in digital archiving in the
    UK which are addressing these challenges, and report on developments in voluntary and legal
    deposit outside the UK.

5. The procedures involved in voluntary or legal deposit are tackled step-by-step. We consider
    how publications will be identified as eligible for deposit – that is, how publishers and
    libraries will decide whether a publication fulfils the exclusion criteria set out in the voluntary
    code, and in particular whether it is seen to be "substantially identical" to a print publication.
    The delivery of publications to a deposit library – whether the British Library or one of the
    five other copyright libraries – is also discussed.

6. We are particularly concerned with the initial registration of publications when they are
    delivered – that is, in the development of the metadata necessary at the point of accession, the
    point at which a protocol needs to be established between the deposit library and the
    depositing publisher. We discuss those aspects of unique identification which impact directly
    on this question, and examine the question of core metadata which publishers will need to
    deposit alongside their publications. We consider the way in which this might be done, and
    conclude that the establishment of standardised processes will be essential to the success of
    the scheme. In turn, a collaborative approach between libraries and publishers will be a
    crucial contribution to establishing such standards.

7. We outline the strategies which the libraries might adopt in preserving their resources:
    refreshment, technology preservation, emulation or migration. We discuss what they will
    need to know about deposited publications in order successfully to preserve resources using
    each of the identified strategies, and what this implies in terms of resources and costs. Finally,
    we discuss the implications of developing catalogue records for electronic publications, and
    of providing reader access to them.

8. The study concludes by pointing out that the original concept of centralised deposit of publications was developed in the wake of a major revolution in technology, the invention of the printing press. We are now experiencing a similar revolution and this is an appropriate moment for a root and branch review of attitudes and processes.

9. It is not possible entirely to divorce technical issues from questions of policy. Technology standards will be critical for all aspects of digital preservation, and a concerted dialogue between libraries and publishers must continue throughout the process of establishing those standards.

Mark Bide
Mark Bide & Associates
mark@markbide.co.uk

Introduction

1.1   Background and scope

Under existing legislation, deriving from the Copyright Act 1911, all books must deposited, free of charge, in the British Library.[1] Under the same act, five other libraries are entitled to receive, on request, *gratis* copies of any book published.[2] This system is known as the "legal deposit" of books.

In 1997, the Secretary of State for Culture, Media and Sport convened a Working Party to discuss the extension of legal deposit to "non-print" materials.[3] Further to the Working Party's report,[4] the Publishers Association and representatives of the copyright libraries are aiming to establish a mutually acceptable code of practice for the voluntary deposit of certain types of electronic materials. It is anticipated that legislation will eventually be framed to cover mandatory deposit of certain classes of non-print materials.

A voluntary code of practice will provide an opportunity for the deposit libraries and publishers in the UK to work through the many issues which arise in the context of the archival deposit of electronic materials.[5] A significant number of these are matters of policy, relating to the terms under which the material is deposited and subsequently accessed. These policy issues are critical, but they are emphatically not the focus of our study. Rather, we concentrate on the technical and procedural standards required to underpin the deposit of certain types of electronic publications.[6]

In order to do this, we draw on the work in progress on digital archiving in the UK and elsewhere, and on the experience of voluntary and legal deposit schemes in other counties. We aim to understand the requirements for standards concerned with the way in which electronic material is held and can be delivered, with particular reference to the metadata associated with the material.

---

[1] Originally, this was of course the British Museum; this perhaps underlines the extent to which the concept of legal deposit may first have been devised for the maintenance of a preserved record of cultural history, although the legislation imposes no explicit *duty* of preservation.

[2] These are known as the "copyright" or "deposit" libraries: the Bodleian Library, Oxford; the University Library, Cambridge; the National Library of Scotland; the Library of Trinity College, Dublin; and the National Library of Wales. We do not know the extent to which the introduction of this distributed system of deposit was consciously introduced as a means of ensuring "preservation through duplication". The more copies of something that are kept (in the physical world, at least) the more likely is one instance of it to survive.

[3] The background to the establishment of the Working Party is given in the Introduction of their Report (see footnote 4). In essence, it was a response to the Secretary of State's concern about the "rapidly increasing gap that is developing in the national collections in respect of non-print material" (paragraph 1.7).

[4] Report of the Working Party on Legal Deposit. This is available at http://www.culture.gov.uk/LDWGRPT.HTM

[5] The voluntary code is in draft form. Comments on the code in this report are based on the Eighth Draft. It is currently unpublished, although we believe that it has been made fairly widely available to interested parties.

[6] It should also be noted that this document does not address the question of the deposit of microform publications (although the code does). See further Section 2.3, "Definitions".

We set out the relevance of existing and emergent standards, and seek to identify any "best practice" that may already have been established.

It is thus our intention to consolidate the findings of work in this area for as broad and 'non-technical' an audience as we can, and, specifically, to encourage thinking along one particular trajectory of the problem, that of standards.

1.2   Method

We began this study by reviewing the literature. We subsequently contacted interested parties and spoke face to face with a number of representatives from the British Library, from relevant eLib projects, and from the publishing industry. Their comments led us to focus on particular projects and studies in this area as the basis from which to progress with our particular interest, the development of standards.[7]

In addition, we developed two questionnaires, one addressed to publishers and one to national libraries in other countries. The first aimed to establish, broadly, the amount of data being published that might be eligible for deposit, as defined in the voluntary code of practice; the formats in which this is currently held; the documentation and metadata currently available to describe that material; and publishers' preparedness, in practical terms, to take part in a voluntary deposit scheme. The second aimed to understand how other national libraries had addressed the technical and procedural aspects of deposit. We solicited responses to the questionnaires largely by telephone interview, although some parties responded in writing. We are extremely grateful for the time given to us by many publishers and librarians, and for their interest in the topic.

This study presents the findings of our research, and our conclusions about the requirements for standards. Literature referred to in the study is given its full bibliographic reference in the footnotes, and is gathered together in Section 6.

1.3   Definitions

There is a large and growing body of literature which relates to what one might call "digital preservation". In this paper, however, we are looking at digital preservation in a very specific context, as set out below.

1.3.1   Types of non-print material

The purpose of the proposed voluntary deposit scheme is to extend the national published archive in the UK by the addition of content from **non-print** publications. Voluntary deposit schemes are already in place for certain "non-print" publications: film, sound recordings, digital mappings.

---

[7] We paid particular attention to studies carried out by DAWG and CEDARS (see further Section 3.4). For the library background, we used as a basis for our thinking the CPA / RLG report, which has had a seminal influence in this area (Waters, D and Garrett, J *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group*. Commission on Preservation and Access, 1996. http://www.rlg.org/ArchTF). In addition, we regularly referred to a report published by the British Library Research and Innovation Centre, which examined the procedures required to support the deposit of one type of offline publication (CD-ROM) in the British Library (see Clarke, A *British Library Legal Deposit CD-ROM Demonstrator Project* May 1997).

The scheme with which we are concerned deals with "non-print publications which are primarily text-based or, if multi-media, which contain a significant level of text content". It covers **microform** publications, **offline** electronic publications and certain types of **online** publications.

Our study does not cover microform publications. By "offline" electronic publications we mean those issued on discrete physical digital media such as tapes, diskettes or, more commonly, optical disks of some kind, such as CD-ROM.[8] Where offline publications contain links to online material, we call them **hybrid** publications.

By "online" electronic publications we mean published resources accessible on the web or on proprietary networks. The voluntary code anticipates the deposit only of those online publications "which are substantially **fixed** at the time of first publication", and does not cover "continuously updated material, such as **'dynamic'** databases". However, in this document we try to address the issues which arise across the spectrum of online publications.[9]

Two points emerge. First, there is not a precise correspondence between the types of material covered by the voluntary code and those covered in this study: while the code covers microform, offline and "fixed" online, we address offline and a range of online publications. We have chosen this approach in line with what we believe to be the areas which most pressingly require analysis. Second, we are focusing only on those activities which together ensure the "preservation" of non-print publications in the context of the national published archive, that is, in the British Library and the other five copyright libraries. While we draw on the experiences of other libraries in digital preservation, and on that of "digital libraries" and "digital archives",[10] in order to examine the issues which the copyright libraries will confront, we are not explicitly concerned with digital preservation as it affects the full spectrum of library services.

1.3.2   Preservation

The glossary of the National Preservation Office defines **preservation**, as it relates to print publications, as "all managerial and financial considerations including accommodation and storage provisions, staffing levels, policies, techniques and methods involved in preserving library and archive materials and the information contained therein".[11] Other definitions similarly stress the *range* of issues which are properly part of a definition of preservation – organisational, technical, financial.[12]

---

[8] They have been variously described in the literature as "offline", "hand-held", "portable" and "packaged" electronic publications.

[9] Our taxonomy of online electronic publications (static, cumulative and dynamic) is discussed in Section 3.3.

[10] We refer the reader to the CPA/RLG report for discussion and definition of these terms.

[11] This definition was used in Foot, M *A Preservation Policy for Digital Material: A Librarian's Point of View* In Fresko (ed) 1996. (Fresko, M (ed), Long Term Preservation of Electronic Materials: A JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib) 27th and 28th November 1995 at the University of Warwick. British Library R&D Report, 6238. 1996.)

[12] Compare, for example, Matthews, G *Preservation Management* LIBS: Library and Information Briefings, 73, August 1997 ("The managerial, financial and technical issues involved in preserving library materials in all formats – and/or their information content – so as to maximize their useful life.") and Conway, Paul *Preservation in the Digital World* 1996 www.clir.org/pubs/reports/conway2 ("The

Where electronic publications are concerned, rather than print, we need again to be aware of the spectrum of activities which enable a library to acquire, organise, make available and preserve its collections. In addition, we need to be more specific about what we mean by "preserving" a resource. Are we aiming to preserve resources so that future researchers can use them – display, search, browse, manipulate them – just as we do now? Or are we rather aiming to preserve all, or some, of the information contained in them? Or elements of the information and elements of the functionality? Preservation activities will differ according to the way in which their *purpose* is defined. These points are critical to a precise definition of "preservation", and we address them further in Section 3.2.4.

We must however be clear about one point. The term "digital preservation" can be used in two substantially different ways. In this study, we use "digital preservation" exclusively to refer to the preservation (whatever exactly that entails) of material which is available solely in electronic form, or "born digital",[13] since this is the concern of the national archive. We do not use the term to refer to the "preservation" of artefactual information by digitising its image (what is sometimes referred to as "retrodigitisation").[14]

1.3.3   Metadata

In this paper, we will use the term "metadata" in its broadest and commonly-used sense of "data that describes things" rather than the more constrained sense of "data that describes data." Purists may object to this broadening of meaning (and to our use of the term metadata as a singular collective noun); however, the amount of circumlocution that would otherwise be necessary would make the task of our readers even more daunting than it already is!

1.3.4   Standards

We take a similar approach to the use of the word "standards". We suggest that the development of standardised processes and adoption of standardised technologies should be an essential target outcome of the voluntary deposit regime. As far as possible, these standards should be based on relevant standards development elsewhere, to avoid unnecessary reinvention and duplication. However, where recognised standards[15] do not meet the need, new developments will be necessary. This does not necessarily imply a formal standardisation process, simply adequate stability to provide the confidence necessary to allow for widespread adoption. Both deposit

---

acquisition, organization and distribution of resources to prevent further deterioration or renew the usability of selected groups of materials.")

[13] The phrase "born digital" has been popularised recently by the Council for Library and Information Resources.

[14] Indeed, the idea that *digitising* printed materials represents a method of *preserving* them may strike the reader as ironic, given the problems that may arise in preserving digitised resources. There is a serious point here: if we do not manage to solve the problems surrounding the long-term preservation of digital products, we now risk losing not only publications "born digital" but also the many resources that have been digitised in the past few years. There is, of course, a significant difference between the preservation of digital resources which have been created with preservation as their primary aim and those which are created for other reasons. What is more, the digitisation of rare (possibly unique) resources will contribute to preservation through duplication – and will allow a degree of access that was previously not possible.

[15] Whether these are *de facto* standards like PDF or *de jure* standards like SGML.

libraries and publishers will need a reasonably stable environment within which to develop the necessary experience.

2   Background: the preservation of digital materials

2.1   *Digital preservation and legal deposit*

Our ability to continue to use digital information is threatened by the limited life of the media on which it is distributed. The National Media Laboratory estimate that the life expectancy of magnetic tape is between two and thirty years, for example, and that of optical media between five and one hundred years.[16] Equally, electronic publications depend on particular software and hardware to access and interpret their data, which become obsolete in the passage of time. As the CPA / RLG report asks, "Who today has a punched card reader, a Dectape drive, or a working copy of Fortran II?"[17] It is also sobering to think that publications which operate under a version of DOS earlier than version 7, or under Windows version 3.1 or earlier, may not be accessible after 31 December 1999.

Given this fragility of digital information, then, we must frame our needs for its preservation. Fundamentally, knowledge – of any type, and in any area – advances through the use and development of previous bodies of knowledge. The advancement of knowledge – which is, after all, what unites all the players in the information chain[18] – requires consistent and reliable access to information sources. Libraries and archives must identify and retain those sources and ensure continuing access to them, both for the scholars of the future and of today.[19]

Clearly, there is a need for digital preservation. The Kenny report[20] concludes that legal deposit represents the best method of ensuring the development of a comprehensive national digital archive. However, voluntary or legal deposit is a very different instantiation of "digital preservation" than that exemplified in other "digital preservation" schemes. In particular, the relationship between the creators of the material (publishers) and its preservers (the deposit libraries) is very different in the national deposit context as compared with most other digital archives, where preservation issues *inform* decisions taken about the way data is created.

For example, the Arts and Humanities Data Service (AHDS) has been influential in its thinking about digital preservation. It recognises that resources that are created in certain formats will be easier to preserve than others, and can therefore suggest as "best practice" that these formats be

---

[16] This data was presented in Keefer, A *Preservation of Electronic Publications* Presentation to the SLA Mediterranean Conference at Barcelona, 26-27 February 1999.

[17] CPA / RLG, Introduction.

[18] As pointed out in the CPA / RLG report.

[19] Indeed, one of the major reasons cited for authors' unwillingness to submit to journals which are available solely in electronic form is their concern that their communication will not remain dependably available.

[20] See footnote 4.

used in its constituent archives.[21] The use of these formats might be more costly if one considers only *data development* costs. However, if one considers the concept of a resource's complete *life-cycle*,[22] and the totality of the costs that spread across it, it is actually the less expensive option, because these formats make preservation easier and therefore less costly.

In contrast, publishers do not undertake the preparation of their resources with a view to long-term preservation, but rather with a view to making the product as attractive as possible to its potential market, aiming to maximise both quality and sales. This may involve the use of formats that will not prove easy to preserve in the long term, and the deposit libraries cannot recommend that the resources are prepared in other, "preservation-friendly" formats instead (or rather, they can make recommendations, but they cannot expect that publishers will necessarily follow them).[23]

An analogy with print preservation is pertinent here. Many publishers have adopted acid-free paper in response to librarians' requests; to do so was more or less without cost to the publisher. They have not also, say, improved the quality of their binding; that too would be an aid to preservation, but it would involve significant cost and would thus not, at first sight, be in the publisher's interest. Extrapolating from this, we should recognise that publishers are unlikely to adopt data preparation standards and practices *purely* because these will assist the deposit libraries, if the costs associated with doing so are significant from the publisher's point of view.[24]

This is not to suggest, however, that publishers see no value in a national archive. Indeed, the negotiations surrounding the voluntary code of practice suggest that they *do* see its value, while being concerned to ensure that it does not damage their commercial interests in publishing their products. Publishers also recognise, as we have mentioned above, that deposit might satisfy one of their direct interests, namely the ability to meet authors' concerns about a permanent archive of their work. More generally, as contributors to the information chain, they are not entirely uninterested in, or unsympathetic to, the development of a national archive which, after all, can prove as fertile a resource for the development of publishing projects as for scholarly research.

---

[21] The archives that constitute the AHDS are: the Archaeology Data Service; the History Data Service; the Oxford Text Archive; the Performing Arts Data Service; and the Visual Arts Data Service. See http://www.ahds.ac.uk

[22] The concept of the "life-cycle" of a resource is central to the Strategic Policy Framework paper (Beagrie, N and Greenstein, D *A Strategic Policy Framework for Creating and Preserving Digital Collections* British Library Research and Innovation Report 107. 1997 The final draft, dated 14 July 1998, is available at http://ahds.ac.uk/public/srg.html. See also Section 2.4.2).

[23] Indeed, the difference in the relationship is recognised in the Beagrie & Greenstein paper: "Data creators who attach little or no importance to the long-term preservation of the data resources they create are unlikely to adopt standards and practices which will facilitate their preservation. This is particularly true when those standards and practices are different from or more costly to implement than those which promise the cost effective development of a data resource capable of fulfilling its intended use" (p4). In the terms of the Beagrie & Greenstein paper, publishers and the deposit libraries are "stakeholders" who are "differently interested".

[24] The result of this – that deposit libraries will be required to preserve resources in a wide range of formats – is also addressed in Section 3.2.4.

2.2   The non-technical issues confronted by digital archiving and deposit

Digital preservation, and in particular legal deposit, raises a wealth of non-technical issues which we can only begin to outline here, and which are not properly our focus. We raise them here to provide readers with a context for the discussion of technical issues, and to give references to papers which address them more fully.

First is the broadly cultural or social issue that we have raised above, the concept of our need for deposit and the importance of what might be called a 'national memory', as embodied in a national archive of published material. It is clearly never possible to archive everything that is "published",[25] any more than it is possible to preserve every historical artefact in a museum. This brings us straight to issues of selection policy: who decides, and how, what we are to archive?[26]

Once this has been decided, there are further policy questions: when and where does deposit take place, and how widely is access made available to deposited publications? These relate closely to the legal and economic issues involved in the management of the deposit infrastructure.[27]

It is important to appreciate at some level at least this wider context in which digital preservation takes place. In this study we focus on the technical issues involved, an understanding of which is essential before the ramifications of those other issues can be fully appreciated.

2.3   The technical issues confronted by digital archiving and deposit

The technical challenges surrounding digital preservation turn on the very concept and definition of an "electronic publication".

The first way in which an electronic publication might be defined is via its delivery medium. In this study, as we have explained, we are interested in offline, online and hybrid offline / online publications.

Second, we might look at the way in which it is "released" as a publication: [28]

---

[25] In this context, a definition of what having the status of a "published document" actually represents in the online environment will be an essential step if a sensible mandatory regime is to be established.

[26] At this stage, the voluntary code aims to cover those publications which most closely equate to the traditional forms of scholarly communication, excluding games, software, and certain other types of electronic publications (see further Section 4.2.1, "Selection"). Interestingly, this excludes one of the types of material that has attracted keen attention in the Internet Archive: election websites. (For more information on Brewster Kahle's Internet Archive project, see Kahle, B, *Archiving the Internet* Scientific American, March 1997, available on the web at http://www.archive.org/sciam_article.html. This article also refers to the use made of archived Presidential Election websites by David Allison of the Smithsonian Institution.) In addition, by excluding games, the code excludes one of the most dynamic areas of electronic "publication", and what may prove to be a significant cultural pointer to the future of interaction between man and machine.

[27] These questions are addressed from a variety of angles in much of the literature. Readers are referred in particular to the CPA/RLG report, to the CEDARS website (http://www.curl.ac.uk/projects.shtml) and to the JISC/NPO studies carried out under the auspices of DAWG (see further Section 3.4).

[28] This taxonomy of resources – static, cumulative and dynamic – and their respective definitions is taken from the analysis adopted by the <indecs> project. "<in*d*ecs> Metadata Model Version 2.0" (1999) Godfrey Rust, Mark Bide. See Section 11.2.17, Creation continuity. The model is available at http://www.indecs.org

- once only, as a resource whose form and/or content is recognised as substantially *fixed* throughout its history; **static resources**;

- cumulatively, as a resource whose content is being *added to* throughout its history (but where the additional content elements are themselves substantially fixed); **cumulative resources**;

- or continuously or 'dynamically', as a resource whose form and/or content is recognised as *changing* throughout its history; **dynamic resources**.

In addition, we must define the content itself delivered via the medium. Electronic publications may consist of:

- data (this might be text, images, video or audio, and will be stored in a wide variety of formats, some standard, others proprietary);

- indexes to the data;

- links to other data;

- metadata;[29]

- software (and we should be aware that that software probably relies on a particular hardware environment and operating system.)

What are the implications for digital preservation? "Preservation" of an electronic publication can only be seen in a very limited sense to mean preserving the medium. Retaining the ability to display, retrieve, manipulate and use the information contained in such a publication, in the face of constantly changing technology, depends rather on preserving all of the items listed above, in line with its release cycle (static, cumulative or dynamic). This is a different challenge altogether. How will a publisher deliver a publication, thus defined, to a library? How will it be identified by both publisher and library? What will it mean for the library to preserve it? What is involved in cataloguing it?

In sections 3.4 and 3.5, we outline a number of projects, in the UK and abroad, which have addressed or are addressing precisely these questions. In Section 4, we work though their implications for each of the processes involved in deposit, bringing to bear the experiences of those projects in seeking solutions to our challenges.

2.4   Overview of current work on digital archiving in the UK

2.4.1   CEDARS

CEDARS (CURL Exemplars in Digital Archives) is a JISC[30]-funded project run by the Consortium of University Research Libraries (CURL). It is concerned with digital preservation as

---

[29] It can justifiably be argued that both indexes and links are metadata; however, we see these types of metadata to be sufficiently distinct to warrant separate inclusion in this list.

[30] JISC is the Joint Information Systems Committee of the Higher Education Funding Councils, the information systems body which works on behalf of the UK higher education and research councils communities. See http://www.jisc.ac.uk

it affects academic and research libraries, its central goal being to "address strategic, methodological and practical issues and provide guidance in best practice for digital preservation". It is undertaking practical demonstrator projects which will provide concrete experience in preserving digital resources using a number of different methodologies.

CEDARS is a three year project due to complete in 2001; its practical case studies will begin in the academic year 1999/2000. One of its key areas of interest in its first year has been the metadata required to describe a resource at the point of its submission into a digital archive.[31] A draft report discussing those requirements is anticipated in late Summer 1999.

For more information, see http://www.curl.ac.uk/projects.shtml

2.4.2   DAWG

DAWG (the Digital Archiving Working Group) was established by the National Preservation Office, largely to oversee the development and publication of seven JISC-funded studies on the preservation of digital materials. Following the publication of those studies, DAWG's official remit has been completed. Proposals for a successor body, or bodies, to DAWG are currently under discussion.

The first publication was an analysis of the CPA/RLG report to identify its relevance to the UK situation.[32] A framework of data types and formats was developed, focusing on the implications for preservation of each,[33] as well as a comparison of possible preservation methods and costing models.[34] Two further studies concentrated on the data creators and the responsibility for archiving,[35] while a further study put this and other issues into the framework of strategic policy, examining how different organisations approach stages in the life-cycle of digital resources.[36]

---

[31] See Day, M, August 1998 *Metadata for Preservation. CEDARS Project Document AIW01* http://www.ukoln.ac.uk/metadata/cedars/AIW01.html

[32] Matthews, G, Poulter, A and Blagg, *E Preservation of Digital Materials Policy and Strategy Issues for the UK: Report of a Meeting on the CPA/RLG Report, December 1996* British Library Research and Innovation Report 41. 1997.

[33] Bennett, J C *A Framework of Data Types and Formats, and Issues Affecting the Long-Term Preservation of Digital Material* British Library Research and Innovation Report 50. 1997.

[34] Hendley, T *Comparison of Methods and Costs of Digital Preservation* British Library Research and Innovation Report 106. 1998.

[35] Haynes, D, Streatfield, D, Jowett, T and Blake, M *Responsibility For Digital Archiving And Long-Term Access To Digital Data* British Library Research and Innovation Report 67. 1997. Also The Data Archive, University of Essex *An Investigation Into The Digital Preservation Needs of Universities And Research Funders: The Future Of Unpublished Research Material* British Library Research and Innovation Report 109. 1998.

[36] Beagrie, N and Greenstein, D *A Strategic Policy Framework for Creating and Preserving Digital Collections* British Library Research and Innovation Report 107. 1997. (The final draft, dated 14 July 1998, is available at http://ahds.ac.uk/public/srg.html.)

Finally, post-hoc rescue or "digital archaeology" was examined.[37] A useful synthesis of the reports has recently been published.[38]

2.4.3   Public Record Office

The University of London Computing Centre (ULCC) acts under contract to the UK's Public Record Office (PRO) as the repository for certain digital resources which it has selected for long-term preservation.[39] The ULCC is primarily responsible for preserving archived data at the bit-stream level, but is also contracted to distribute the data to secondary users (both on physical media and online). Currently, archiving and preservation is handled by storing multiple copies off-site and on-site, and by the periodic refreshment of data to new media.[40]

2.5   *Overview of developments in legal deposit and digital archiving outside the UK*

Deposit operates under a variety of regimes in other countries. Print, microform and offline electronic publications are covered by voluntary and legal schemes, but there are few mature schemes which cover online publications. Many of those that do exist are focused at present largely on university dissertations and theses.

Sweden, Denmark and Finland have legal deposit systems which cover offline and online materials. Denmark's online policy only covers those online publications which are "complete", and their definition of "complete" excludes electronic newspapers and periodicals. Finland uses automated data harvesting to archive publications that are freely available on the web, but publishers are obliged to deposit online publications that are not made freely available.

The Library of Congress in America has mandatory deposit for certain types of print, microform and offline publications. Its Copyright Office, however, which accepts publications for copyright registration, covers print, microform, offline and online publications, although less than a thousand online registrations have been made to date.[41] The National Library of Canada operates legal deposit for offline publications, and is looking into the extension of legal deposit to online publications. For the time being, these are collected on the basis of individual negotiation with publishers, and the Library has collected just over two thousand titles to date (some of them serials).[42]

---

[37] Ross, S and Gow, A *Digital Archaeology: The Recovery of Digital Materials At Risk* British Library Research and Innovation Report 108. 1999. This study provides an object lesson the importance of developing digital preservation strategies that are *not* "post-hoc": while post-hoc rescue can produce results, it is extremely costly and laborious.

[38] Feeney, M *Digital Culture: Maximising The Nation's Investment. A Synthesis of the JISC/NPO Studies on the Preservation of Electronic Materials.* The National Preservation Office: 1999.

[39] There is a case study in Beagrie & Greenstein 1998, p20ff. (The paper is summarised in Feeney 1999, Chapter 3.)

[40] On refreshment and other methods of preservation, see further Section 3.2.4.

[41] Online registrations are handled in the CORDS system: see http://www.loc.gov/copyright/cords for more information. Most of the registrations to date have been dissertations, through agreement with UMI (see http://www.loc.gov/today/pr/1999/99-007.html).

[42] For general information, see http://www.nlc-bnc.ca

Similarly, legal deposit in Die Deutsche Bibliothek and the Bibliothèque nationale de France covers print and offline publications, and hybrid publications are treated as if they were offline (that is, the online component is effectively ignored).[43] In addition, Die Deutsche Bibliothek accepts online dissertations from certain universities and operates a voluntary scheme for the deposit of electronic publications with selected publishers.

The Koninklijke Bibliotheek in the Netherlands, on the other hand, operates a voluntary scheme for all print, microform and offline publications, as well as some online (dissertations and journals). They are not seeking legal enforcement for this system, which works well on a voluntary basis. The National Library of Australia also runs a voluntary scheme for the deposit of offline resources; they are anticipating that they will move to legislation.[44]

Each of the libraries is developing policies and systems for preservation, and we detail their experiences in the relevant parts of Section 3. However, the responses to our questionnaire consistently stressed that there is an enormous amount of research which needs to be completed before many aspects of policy and practice can satisfactorily be resolved.[45] Although there is a growing body of expertise in this area, experience in digital preservation is still in its infancy.[46]

In seeking solutions in certain areas, many countries are collaborating on two major projects which are important to consider here.

NEDLIB (Networked European Deposit Library) is an EC-funded project which aims to construct a basic infrastructure which could be implemented by its participating partners to handle the electronic collections they are developing through deposit. Partners include the national libraries of the Netherlands, France, Norway, Portugal, Switzerland, Italy and Germany as well as a number of other libraries and three publishing companies.[47]

NEDLIB aims explicitly to ensure that electronic publications of the present can be used now and in the future. It plans to deliver a "toolbox" for the operational management of electronic deposit systems, containing three modules (covering accession and installation, access and archiving), and capable of being embedded in existing IT infrastructures in national libraries. At present, the project is developing a proof-of-concept demonstrator. It has not yet developed a policy on access to digitised resources.

For more information, see http://www.konbib.nl/nedlib

BIBLINK is an EC-sponsored project led by the British Library which involves the national libraries of France, the Netherlands, Norway and Spain, as well as UKOLN at the University of

---

[43] See http://deposit.ddb.de/index_e.htm and www.bnf.fr

[44] See Cathro, W S *Digital Libraries: a National Library Perspective* January 1999. Documents relating to the voluntary deposit of offline publications are available at http://www.nla.gov.au/policy/vdelec.html

[45] Compare a recent survey of its member libraries conducted by the RLG, which concluded that, where preservation is attempted, it is rarely done so systematically, and that more than half the total number (54) of responding institutions could not access some of their material, due to lack of operational or technical capacity to mount, read of access the data. See Keefer 1999 (op cit).

[46] In this we agree with the opinion defended in the CPA / RLG report (see note 4), whilst acknowledging, as the report does, that expertise has been developing in "digital preservation" for many years.

[47] Although Britain is not directly involved in the project, it keeps a watching brief.

Bath (UK) and the Universitat Oberta de Catalunya (Spain). The project's focus is bibliographic metadata: it seeks to establish links between national bibliographic agencies (NBAs) and publishers of electronic documents for the purpose of exchanging bibliographic data.

A demonstrator is currently testing the proposed workflow, in which publishers would send bibliographic metadata to the BIBLINK "workspace" for the NBA to download, enhance, include it in its national bibliography and finally return it to the publisher to use in their own resources and systems. The project has successfully tested the first stage of data transmission: data can be sent via email or a web form to the workspace (hosted at UKOLN) in a format based on Dublin Core and extended for library purposes, which is known to the project as "BIBLINK Core". The British Library then coverts the data to UNIMARC, which the participating NBAs in turn convert to the appropriate national MARC format and enhance (for example, by applying name authority control and subject classification).

The project is currently testing the next stage of transmission, in which the NBA returns the enhanced, authoritative record to the workspace for the publisher to download and use. A demonstration of this is anticipated very shortly.

For more information, see http://hosted.ukoln.ac.uk/biblink

3   Deposit in the UK

3.1   The aim of deposit

On the basis of what we have learned, we believe that it will be a primary task for the UK's deposit libraries, during the voluntary deposit phase, to determine the long-term function and purpose of the digital collections that they are developing. This will in turn determine the strategies they adopt for their organisation and preservation.

As we understand it, the general purpose of deposit, as discussed in Section 2.1, is to ensure that the knowledge of the present is not lost to the future. This does not necessarily imply that deposited publications will be made available "anywhere and anytime" for future researchers (or indeed for current researchers). The aim of deposit is not to make resources *widely* available, but rather to ensure that they are not lost, and are accessible for future consultation *in however limited a context*.[48]

At one end of the scale, deposit libraries might be aiming to provide access to all the electronic publications deposited with them to their users in twenty or fifty or five hundred years' time. These users would see exactly what we see now when we access the resource, and would be able to perform exactly the functions we perform – browsing, searching, downloading.

For reasons which will become clear, this scenario appears unlikely to be possible with the present generation of digital resources, particularly those distributed on offline media.[49] Librarians in national libraries whose deposit systems have included electronic publications for

---

[48] The issue of access to deposited publications is discussed further in Section 3.2.6.

[49] It is certainly arguable that the problems of digital preservation may reduce over time if modes of publication become more standardised in response to market demand.

some years and to whom we spoke seem fully resigned to this: indeed, already they cannot access a significant proportion of the material deposited with them in the 1980s.[50]

Another scenario might envisage the user of the 2020s, say, or the 2050s, being able to use the *information* contained in some of the deposited publications, but perhaps visual aspects of their presentation might have substantially changed, and perhaps elements of their functionality might have been lost. In a development of this scenario, curators might have used the advances of technology to *improve* aspects of the resource – to sharpen photographic images, for instance.[51] We need to decide what we are *aiming to achieve* in our digital preservation, and, if preservation cannot be achieved without losing certain aspects of the original, to what extent it is acceptable to change or modify it (including, for example, adding functionality).

Clearly, these questions raise many issues – and possibly a few hackles. What is implied for the integrity and authenticity of publications? We cannot back away from addressing these questions, for the inevitable corollary would be that users of the future would not be able to access the resources of today *at all*.

3.2   Procedural issues

In this section we attempt to tackle the ways in which libraries and publishers will come up against the questions raised above, working through each procedural stage involved in deposit. We follow each stage of the library's interaction with deposited material, attempting to answer what the deposit libraries need to know about a publication in order to accession, preserve and catalogue it, and the extent to which the publishing community can currently provide that information. We try to draw out what this implies about standards developments required.

In writing this Section, we have made particular reference to two papers which have started to address the requirements of the British Library in dealing with deposited material: the consultancy study carried out by Cimtech and presented at the Warwick workshop in 1995, and a study co-ordinated by British Library Research and Innovation Centre which investigated the readiness of the Library to receive offline publications.[52] We emphasise that these studies were used to gain an appreciation of library processes and issues, rather than suggesting that they have laid down a blueprint for actual procedural policy for either the British Library itself or for any of the other deposit libraries.

---

[50] This raises questions which are absolutely central to discussions concerning deposit; if its aim is not long-term preservation, what is its justification?

[51] This was a point raised in Arms 1996 in relation to retrodigitisation. See Arms, C R *Historical Collections for the National Digital Library. Lessons and Challenges at the Library of Congress* D-Lib Magazine April 1996 (p5). Available at http://www.dlib.org/dlib/april96/loc/04c-arms.html

[52] See Hendley, T *Practical Implications: Long-Term Preservation of Electronic Materials* In Fresko (ed) 1996 (op cit) and Clarke 1997 (op cit). In fact, the Demonstrator project documented in Clarke 1997 addressed readiness to receive CD-ROMs only. For our purposes, this represents a limited subset of material, but many of the lessons learned are nevertheless pertinent here.

3.2.1   Selection

Whereas for printed materials, the publisher of every book published in the UK is required to deposit one copy in the British Library,[53] the intention is that only *certain* non-print materials will be required for deposit (as explained below). Initial "selection" procedures will therefore have an additional dimension in the context of the deposit of non-print materials.

3.2.1.1   Comparison with print

The voluntary code of practice states that "publications which appear with substantially identical content in more than one medium only need to be deposited in one medium", and that, for the time being, "the choice of the deposit libraries would normally be print where that was available". This goes for both offline and online publications.[54] A process is required whereby publishers and/or the libraries determine in which media publications will be deposited, according to an agreed understanding and definition of what it means for content to be "substantially identical". At present, there is unlikely to be any indication from the way an electronic product is identified that it has essentially identical content to a related print product.[55]

Further, this must be an iterative process, rather than a once-only process, for those publications whose content is released or updated on a regular basis (that is, for cumulative and dynamic resources, as defined in Section 3.3). For example, a publisher might publish an online version of an encyclopaedia which in 1999 is substantially identical to a print product. However, the online version might be developed such that by, say, 2003 it contained substantial additional content when compared with the printed publication, and would therefore become eligible for deposit.

Comparing offline and online publications with a print product, or a number of print products,[56] is conceptually simple although it would be laborious in practice. In the case of hybrid products, both the offline and the online components need to be considered. A strategy will have to be formulated to meet those (frequent) cases in which the offline component is substantially identical to a print product, but the online component is not.[57] Would there be any value in archiving the online content without the offline?[58]

---

[53] See the Copyright Act, 1911. Interestingly, it seems that certain definitions were taken for granted (like what a "book" is).

[54] Eighth draft, p4. On p5, the wording is that publications would be deposited "unless they are substantially identical in content *and functionality* to print publications" [our emphasis]. This wording gives us some cause for concern. The idea that an electronic publication might have essentially the same *functionality* as a print on paper version of the same content (as opposed to fulfilling the same *function*) seems inherently unlikely. This ambiguity needs attention.

[55] On the identification of electronic publications, see further below, Section 4.2.3.

[56] A number of publishers we spoke to are working on online publications whose content they consider to be "substantially identical" to a *number* of print products.

[57] By their nature, online components of hybrids are unlikely ever to be static; they will almost always be either cumulative or dynamic.

[58] However, there would be technical problems associated with preserving the links between the two, as we discuss in 4.2.4.

Our conversations with publishers revealed that there is very little content currently being published that is available solely in electronic form which meets the selection criteria.[59] For example, most scholarly journals that are published electronically currently have some kind of print equivalent – either a "direct equivalent" (in one publisher's terms) or a print publication which has an equivalent "core" or "centre" (in several other publishers' terms). Nor is there much other (ie, non-journal) scholarly content published solely in electronic form which meets the selection criteria. This can be illustrated by looking at the *Waterlow New Media Product Database* (derived from the *TFPL Multimedia and CD-ROM Directory*).[60]

However, electronic publication is growing year on year, and publishers have many plans for the development of "electronic only" content within the next five years – entire publications without a print equivalent, on the one hand, and, on the other, many more electronic-only "add-ons", particularly in scientific journals.[61]

3.2.1.2   Other exceptions

There are further exceptions to the requirement to deposit: offline and online publications are excluded from the voluntary code if they are computer software or games, if they are produced by or for organisations for their private internal use only, if they are already supplied to a legal deposit library through a publishing agreement, or until twelve sales of the publication had been made. Further, online publications are as yet not to be deposited if they are 'dynamic'.[62]

3.2.1.3   Selection practicalities

In checking for deposit eligibility, the libraries will come up against a two-fold problem: first, electronic publications are not "registered" anywhere, which means that a complete listing of all UK electronic publications does not exist.[63] Nor is there a standard identifier that is widely used to identify electronic publications (as the ISBN and ISSN are used for print publications),[64] and which could be used to compile such a listing.

Secondly, there is nothing in the way an electronic product is identified which would consistently indicate its relationship to a printed product, or that it is software or a game, or that it has been produced for an organisation's internal use, or that it is otherwise ineligible for deposit. Selecting for deposit on a publication-by-publication basis by deposit libraries could become an intensely laborious process. Alternatively, publishers might be expected to be the sole judges of which of

---

[59] For this reason, we found it impossible to "size" how much material would be deposited under a strict interpretation of the Code of Practice – but in reality it would probably be very little if the two criteria of "fixed but not substantially different from print product" were strictly adhered to.

[60] See http://www.newmediainfo.com

[61] An interesting related issue is whether or not the electronic-only "add-on" is considered part of the validated paper – and indeed whether or not it is peer-reviewed.

[62] We assume, although it is not clear from the current draft of the code, that our use of "dynamic" and that used by the authors of the code is essentially the same, and that their term "static" covers resources that we would call either "static" or "cumulative".

[63] The most comprehensive listing available is the *Waterlow New Media Product Database*. This contains information on CD-ROM, DVD and other new media titles which are commercially available worldwide. See http://www.newmediainfo.com

[64] On identification, see further Section 3.2.3.1below.

their electronic publications meet the deposit criteria. The most effective mechanism for making this judgement will need to be explored during the operation of the voluntary deposit regime.

3.2.2   Delivery of publications to a deposit library

Currently, one copy of a book is delivered to the British Library, and a demand can be made by the Copyright Agent in London on behalf of the five copyright libraries for up to five copies (one each) of a printed publication to be deposited. The Agent is then responsible for dispatching the publications to the relevant libraries.[65] The libraries choose which publications to demand by publisher, since they would regard it as an unacceptable drain on resources to operate selection at any lower level (such as on a title basis).

The situation envisaged is different, however, for non-print publications. Here, the voluntary code, (which, again, we should stress is only a trial code of practice, and not a blueprint for legislation), suggests that one copy of an electronic publication would be deposited, *either* with the British Library *or* with the Copyright Agent for transfer to the most appropriate legal deposit library. The other libraries could each request an additional copy, but the publisher would not be obliged to comply with any such request. Publishers are expected to indicate on "a standard form, submitted with each deposited publication", whether or not they authorise the deposit library to provide access over a secure closed network to each of the other deposit libraries.[66]

But what does it actually mean for a publisher to deposit, and for a library to receive, an electronic publication?

In the case of offline products, delivery is a fairly simple concept, since diskettes, CD-ROMs and other portable media can be packed and sent like a book. In fact, some publishers are already delivering electronic publications to the Copyright Agent as a matter of course, alongside their printed works. The deposit libraries are making available some of these publications, where resources and technical capacity allow; others are being retained, and preserved on a "best efforts" basis, until those issues can be resolved.[67]

In the case of the online component of a hybrid product, or a fixed online publication itself, the situation is more complicated. To give a library a 'copy' of the publication involves making available a copy of everything that makes it functional, that is, all the data, indexes, links, metadata, software which constitute it (as discussed in Section 3.3), which must then be run on the required hardware specification. [68] Only then would the library be able to host the publication on its own servers, as a "replica" of the publisher-hosted resource.

---

[65] See the Copyright Act, 1911.

[66] Eighth draft, p6 (for offline publications) and p7 (for online).

[67] We are not aware of any problems that may have resulted from the ad hoc approach that is currently being taken to access and preservation issues. This does not mean that they do not exist.

[68] Of course, this assumes that publishers have the right to deposit all these items, including the necessary code and scripting. This is a large assumption to make, but it is beyond the scope of this study to discuss rights and legal issues more fully. A brief discussion can be found in Section 3.2.4.5.

In this connection, it is important to appreciate the difference between 'active' and 'passive'[69] sites. 'Passive' sites are those in which the data exists on the site in a format such as HTML or PDF. 'Active' sites, by contrast, allow users to query a database which generates data "on the fly", or "dynamically", in response to a user request. In the case of passive sites, publishers could deposit data files, but in the case of active sites they would need to supply the database and all the scripting required to generate the data as the user sees it.

It should be noted that this situation implies the exchange of numerous large files, and the use of a great deal of human resource. In addition, problems will arise as the libraries begin to host multiple publications which might conflict with each other, for example by running on different versions of the same software.

Deposit libraries in other countries whose policies cover the deposit of online material receive some of their online material via protocols such as HTTP or FTP. [70] Alternatively, files might be transferred to an offline medium for delivery.[71]

Beyond these issues, the locus of deposit – to any one of the six libraries, with the possibility (which the publisher can deny) of networked access to the other five – raises a point which bears highlighting. It implies that the national archive will operate in a distributed framework, meaning that the expertise required to manage all of the preservation processes described here will need to be developed in each of the six libraries. The distribution of such sophisticated technical and managerial skills suggests a significant investment on the part of each of the libraries.[72] It seems to us that this is a policy which should be carefully reviewed if considerable duplication of effort and investment are to be avoided.

3.2.3 Accession

Here we are concerned with aspects of accessioning which occur in the libraries when materials are first deposited, before consideration is given to preservation issues and before a full catalogue record for the publication is compiled (which is covered below in Section 4.2.5, "Record

---

[69] These are sometimes referred to as 'static' sites, but in this study we will call them 'passive', since we use 'static' in a different way (see above, Section 2.3).

[70] In considering protocols for the delivery of online content, we suggest the investigation of the Information and Content Exchange (ICE) Protocol, for its apparent automation of content and metadata exchange on the Web. See http://www.w3.org/TR/NOTE-ice However, it is worth noting that, despite the impressive credentials of some of its backers, this protocol remains a "W3C Submission" and may therefore not progress towards the W3C RFC status that is the closest that the World Wide Web community has to a formal standard.

[71] Certainly, those publishers to whom we have spoken about this issue have found that the distribution of very large databases can only be realistically effected by tape transfer.

[72] We emphasise here that we are talking about the locus of *initial deposit*, and therefore of *preservation*. The suggestion is that resources be deposited across the six libraries (as outlined above); this will imply that each library execute the refreshment, emulation or migration required for those resources (see further Section 4.2.4). We are *not* talking here about distribution of *access* to the resources once they have been deposited – that is another matter entirely and falls largely outside the scope of this report (see Section 3.2.6).

Creation").[73] These are the processes with which we are particularly interested, because they represent the point at which the standard and effective transfer of information from publisher to librarian will be crucial. It is at this point that we need to appreciate the importance of unique identification.

3.2.3.1   Unique identification
The challenges associated with the unique identification of digital content have been well rehearsed elsewhere[74] and we will not attempt a detailed discussion of all the issues in this document. However, there are some aspects of the topic which have a direct impact on the question of the deposit of digital manifestations and which therefore deserve some consideration here.

Deposit has conventionally involved consideration only of "manifestations" of works.[75] These physical manifestations – particularly books – have been susceptible to relatively straightforward unique identification. For over a quarter of a century, almost all the books deposited at the British Library will have had an ISBN, a way of identifying a specific manifestation. Indeed, the ISBN has been used as one of the mechanisms for identifying products that should have been deposited but have not.

The identification of periodical publications is somewhat more complex, since the ISSN is in reality a *work* identifier; the introduction of standardised unique identification of individual issues (and articles within issues) is much more recent, following the introduction of the Serials Item and Contribution Identifier (SICI). Nevertheless, a standard, unique identifier is widely used that can (for example) discriminate between two serial publications with the same name.

With printed publications, it is thus possible for relatively unambiguous communication to take place between library and publisher. As we move in the direction of electronic products, particularly online publications, the situation becomes more complex.[76]

It is broadly true that offline digital products, particularly those which are included in the terms of the deposit scheme, follow much the same pattern as printed publications in terms of

---

[73] Terminology differs slightly in covering these processes (see for example Clarke 1997 compared with Hendley 1995). We have settled on "accession" here in line with advice from library colleagues.

[74] See, for example, Green, B & Bide, M *Unique Identifiers – a brief introduction* (revised version March 1997) London: Book Industry Communication. http://www.bic.org.uk/bic/uniquid.html; Paskin, N *Information Identifiers* Learned Publishing 1997: Vol 10 No 2 http://www.elsevier.nl/locate/infoident; and Paskin, N *DOI: Current status and outlook* D-Lib 1999: Vol 5 No 5 http://www.dlib.org/dlib/may99/05paskin.html

[75] Although cataloguing theory has been much exercised over the matter of underlying works.

[76] However, it is perhaps worth pointing out here that there may, in fact, be considerable ambiguity about the identification of physical manifestations. For example, the point at which a "corrected reprint" (with the same ISBN) becomes a "new edition" (with a new ISBN) is entirely undefined and will differ between different publishers. This is a clear case of what the <indecs> project refers to as "functional granularity". To paraphrase, something needs to be uniquely identified when it is necessary to differentiate it from something else. The decision of when it is *necessary* is a matter of judgement.

identification. They are almost certain to have either a print-related identifier like an ISBN or a Unique Product Code/EAN barcode.[77]

Matters become further complicated with online products. Things are not made any easier by the decision of the ISSN Agency to mandate the use of a different ISSN for the electronic version of a printed serial publication (even if the two are fundamentally and logically identical). While it may often make sense for different manifestations of the same intellectual content[78] to have different identifiers, we would argue that the underlying abstract work, the journal itself, is the same. This makes decisions on what is and is not identical in intellectual content even more difficult to discern.

There is, as yet, no equivalent of the ISBN for the 21st Century, although this is a position for which the Digital Object Identifier is enthusiastically bidding.[79] Certain genres of online electronic product may in the next year or two be routinely identified with DOIs, particularly online journals where DOI adoption seems to be achieving significant market penetration among the major players. However, we would expect it to take a number of years for the DOI (or another URN) to be used universally in any sector.

The question of unique identification of digital resources is a particular – and very important – instance of the problem of metadata which we will address further below. Once inside the deposit system, it is arguable that the matter of unique identification becomes one for the libraries alone (in the sense that in a closed universe, which we expect the deposit system to be, unique identification is a matter only for those who manage the system). Nevertheless, we would strongly recommend that careful thought be given to the ways in which unique identifiers are used at the point of deposit, particularly to the extent that they may be used in the future as finding and location aids.[80] The long-term value of the deposit archive will depend, to some extent at least, on the approach taken to identification.

3.2.3.2   Metadata at the point of accession
A great deal has been written about metadata requirements for preservation.[81] We will briefly discuss these in the next section. However, as we have already pointed out, the extent to which

---

[77] However, this identification is used for both offline and hybrid publications. In other words, the identification does not make clear to its recipient whether it identifies an offline product, or the offline component of a hybrid product. This is important for the deposit libraries to know, if they wish to preserve the online component of the hybrid.

[78] It is important to make clear that content night not be only text, but a combination of text and graphics; if the "graphics" are "dynamic" (for example, a revolving spine in a paper on surgery) there can thus be content which is not "shown" in the printed form.

[79] See Paskin 1999 (op cit).

[80] It might well be argued that a digital deposit copy should be given a unique *instance* identifier – probably a library-issued DOI. We can find several reasons to support such an idea, not the least of which is that (assuming that preservation is aimed at the content not the product – see further Section 4.2.4) within a short period of time the preserved instance might have little apparent in common with the manifestation as deposited in terms of "look and feel" or functionality. There might also be some positive security advantages from such an approach.

[81] See in particular: Bearman, D *Reality and Chimeras in the Preservation of Electronic Records* D-Lib Magazine April 1999. http://www.dlib.org/dlib/april99/bearman/04bearman.html; Bearman, D and Sochats, K. *Metadata Requirements for Evidence* 1995 http://www.lis.pitt.edu/nhprc/BACartic.html; Day 1998 (op

there need to be external standards for *preservation* metadata will depend to a great extent on how "open" the access is to the preserved content. Clearly *internal* standards are essential and there may be much to be gained in the long run from learning from the experience of others.

However, we are rather more concerned with the metadata necessary at the point of accession, since this is the point at which there needs to be a protocol established between the deposit library and the depositing publisher. If metadata is to be supplied from publisher to library, standards for formats and delivery protocols will essentially need to be developed.[82]

Metadata for deposit copies of printed media can be catalogued in the traditional way – "book in hand". The primary metadata is available to the cataloguer from straightforward inspection of the printed product. The publisher has no direct influence on the way in which the book is described in the library catalogue.

This is much more rarely the case with even offline digital products. As discussed above, a unique product identifier is not yet in "standard" use, and it can even be difficult to determine the publisher[83] and title[84] of a publication. (Indeed, even if the publisher can be identified, the form in which it is given is notoriously unstable.)[85] Librarians we spoke to expressed frustration at the variety of (non-standard) ways in which publishers currently display and express this critical information.

This is entirely without consideration of the explosion in the number of "publishers" which has accompanied the growth in the World Wide Web. The definition of a publisher in the online environment will prove just as uncertain as the definition of a publication (see section 2.3). In the

---

cit) and Rothenberg, J., *Metadata to Support Data Quality and Longevity*. Proceedings of the 1st IEEE Metadata Conference, 16-18 April 1996.
http://www.computer.org/conferen/meta96/rothenberg_paper/ieee.data-quality.html

[82] In all discussions of metadata standards, it is important to remember that publishers have not traditionally been closely involved in the management of metadata for their publications. Although the importance of metadata is increasingly widely recognised in the publishing industry, there is a considerable way to go in both system and process development before all publishers will be in a position to deliver consistent and accurate metadata electronically.

[83] As discovered, for example, in the CD-ROM Demonstrator project (see Clarke 1997, paragraph 17).

[84] See, for example, Clarke 1997, paragraph 11. If the product identifier was *also* unclear, this could lead to an inadequate record. When a CD-ROM was loaded, however, it was usually possible to find one or the other. (It should be noted that the project tried to register CD-ROMs *without* loading them as far as possible: this makes time efficiencies, and means that staff do not have to deal with the effects of multiple installations on their workstations – and do not have to check for viruses.)

[85] There are two problems: first, there is an enormous variation in the way in which the same publisher or imprint name is given across a number of different publications, although publishers do make attempts (with varying degrees of success) to standardise name forms editorially. In addition, these names change frequently, as a result of company changes, re-structures and mergers. Combined, these two factors cause considerable difficulties for librarians attempting to register and catalogue their collections. A current BNBRF project is in progress, looking at the feasibility of introducing a standard numbering system for publisher and imprint names. If introduced, this would greatly simplify one element of the metadata required on deposit. A report on this project will shortly be available; in the meantime, information can be supplied by contacting Brian Green on brian@bic.org.uk

same way that publishers may have to define which of their products meet the criteria for deposit, may they also have to decide whether or not they are, in fact, publishers at all?

Even if deposit were primarily to be aimed at offline media,[86] there is patently a need for publishers to deposit certain core metadata alongside the content.[87] It is possible to imagine that this metadata might be deposited on a hand-written form[88] although the irony of such an approach must be apparent to all our readers. An alternative approach might be for publishers to ensure that the necessary metadata elements (like unique identifier, title, author(s), publisher) are clearly and unambiguously displayed in the packaging or other literature which accompanies the product. This would allow for effective "disk-in-hand" cataloguing (without all the issues attendant on "disk-in-machine" cataloguing).[89]

The *quality* of the information produced on such packaging is an issue that is worth touching on. Metadata printed on the jacket of a book (information about imprints, title, series title, author) often differs from the information printed on the title and title verso.[90] With an offline publication such as a CD-ROM there is a similar problem, in that the object itself may be handled by a different department from the one handling the packaging; indeed, the problem is likely to be more acute because the person producing the packaging may never have seen any aspect of the content of the CD-ROM.

Even with offline products, then, much remains to be resolved. With online products, the problem becomes more acute.

We have already discussed some of the issues relating to assignment of unique identifiers to digital manifestations of content. The problems relating to other types of metadata are at least equally taxing. Without a "book-in-hand" or even a "disk-in-hand" to catalogue, there are only two possible approaches: either the essential metadata must be extracted from the product itself in conventional ways – likely to prove very difficult indeed with online resources – or the publisher must deposit the metadata alongside the resource being deposited.

In this connection, we find the approach being taken by the International DOI Foundation persuasive (see Paskin 1999). This work, which is closely related to the work of the <indecs> project,[91] proposes that a limited *kernel* of metadata should be deposited alongside every

---

[86] A pragmatic if somewhat short-sighted solution, since it seems to us that offline media are likely to be a relatively transient phenomenon.

[87] Much thought is being given to the necessary elements included in this metadata. Biblink specifies 18 fields in Dublin Core format, for example (see http://hosted.ukoln.ac.uk/biblink). Information is also available about the "legal registration form" that French publishers are required to complete on deposit (see http://www.bnf.fr). Die Deutsche Bibliothek, under the auspices of the German Publishers' Committee, are undertaking a survey to address this issue for online publications, but have already determined the metadata required for dissertations (see http://deposit.ddb.de/index_e.htm).

[88] This is the approach of the Bibliothèque nationale de France, for example (see the previous note), and the National Library of Canada (in the case of its offline publications, which are deposited by law).

[89] Having said this, we would seriously question whether a deposit system which does *not* involve loading and assessing digital content at or near the point of accession has much to do with long-term preservation.

[90] This is a manifestation of the well-recognised problem that many publishers experience in the management of data, particularly where the same data is needed by more than one application.

[91] http://www.indecs.org

registered DOI. The kernel metadata will be supplemented and qualified for different *genres* of content.[92] An approach similar to this, in which a minimal, but tightly defined set of metadata is expected to be deposited by the publisher, would appear to us to be a realistic approach.

The most crucial contribution to success will come from a collaborative approach, between libraries and publishers, to the establishment of the metadata schemas to be used for different *genres* of product. We would suggest that these should be built on the basis of a coherent logical model rather than the somewhat minimalist discovery-only approach taken by (for example) Dublin Core.[93]

There is good precedent for the adoption of a relational approach to bibliographic records in the work of the IFLA Study Group for Functional Requirements for Bibliographic Records (FRBR) (IFLA 1998).[94] This work has been closely studied in the development of the <indecs> logical model, with which both the DOI metadata structure and the recently announced EPICS[95] data dictionary for the book trade are compliant. The closer any deposit metadata standard is to other models being adopted within the publishing industry, the more likely it is that it can and will be complied with.

3.2.4   Preservation

Consideration of preservation issues is placed here in recognition of the fact that libraries need to consider their strategy for the preservation of an electronic publication early in its life. Where one might be able to put off the *processes* required to preserve a print publication for some time after acquisition, the rapid evolution of software and hardware demands that librarians attend to preservation issues at an earlier stage in an electronic publication's life.[96]

This is not to suggest, however, that preservation is a once-off process which falls easily into a single 'slot' in the deposit process.[97] Ensuring long-term access to digital material requires an *ongoing* commitment to refresh or migrate data. Planning those activities is made even more difficult because of the impossibility of predicting future changes in technology.

---

[92] In this context, a genre is a somewhat arbitrary but nevertheless entirely comprehensible classification of a piece of content: for example, a journal article or a book.

[93] BIBLINK has taken an approach built largely on Dublin Core principles.

[94] IFLA Study Group (1998) *Functional Requirements for Bibliographic Records* K G Saur. See http://www.ifla.org/VII/s13/frbr/frbr.pdf

[95] EDItEUR Product Information Communication Standards. The second version of EPICS documentation will shortly be available; in the meantime, information can be supplied by contacting Brian Green on brian@bic.org.uk

[96] This point is made, for example, by Graham, P *Preserving the Digital Library* and Foot, M *A Preservation Policy for Digital Material: A Librarian's Point of View* (both in Fresko (ed) 1996, op cit) and in Keefer 1996 (op cit).

[97] Again, this contrasts with the preservation of printed material, as Foot points out: "In contrast, once a book has been conserved, one can be reasonably satisfied of its continued existence (provided the item is properly stored and not over-handled). Similarly, once one has made a microfilm, provided the film and its production methods are of archival quality and it is stored in the right conditions, the contents of a book or manuscript will be preserved for about 300 years." (See Foot 1995: reference given in previous note.)

The difficulties associated with preserving for the long term have led to Rothenberg's ironic statement: "Digital information lasts forever – or five years, whichever comes first".[98] Others insist that the preservation of information in a digital form is an entirely unfeasible aim in itself, and that the best option is to print out data on acid-free paper.[99] Certainly, the problems are so frustrating that is worth emphasising the unique benefits that digital preservation might bring: the continued ability to browse, search, download and otherwise use a panoply of resources, in ways that are simply impossible with paper or non-digital formats.[100]

If, then, we are to create a national digital archive, policies, techniques and methods need to be evolved which will fulfil the central requirement of preservation: the ability to *access* the preserved resource in the future. There are a number of strategies for preservation, which we discuss below. Finally in this section (3.2.4.6), we discuss the metadata research agenda implied by those strategies.

3.2.4.1   Technology preservation
Using this preservation strategy, the data is maintained along with the original hardware and software on which it depends. This would enable future users to view data and functionality exactly as they were originally published. However, it would effectively create thousands of "museum pieces". Clarke concludes that this seems "not a viable proposition except in very exceptional circumstances".[101]

However, we believe that we should be careful not to write off this strategy too quickly. While it may not be a realistic option in the long run, and would be difficult to manage on any significant scale, it could prove useful for the deposit libraries in the short to medium term (especially given the fact that other, more robust strategies are still being tested and have uncertain outcome). It may be that the most feasible way (technically and economically) of ensuring preserved access to a *limited sample* of our present PC-based applications for users in 2020 *will* be to preserve a number of "museum" machines in our deposit libraries.

3.2.4.2   Refreshment
In refreshment, a complete data resource is simply copied from one physical medium to another. This is the solution of choice for the short term if it is simply the medium (such as the CD), rather than the hardware or the software, that is considered likely to fail.[102]

---

[98] Rothenberg, J. *Ensuring the Longevity of Digital Documents* Scientific American Vol. 272, Number 1. January 1995.

[99] Michael Gorman, for example, argues strongly that "the only practical manner in which we can preserve our present for posterity is to create print on paper archives", since digital archives "are confronted with insuperable economic, technological and practical problems" (See Gorman, M. *What is the future of cataloguing and cataloguers?* Paper presented at the IFLA General Conference, 1997. http://ifla.inist.fr/IV/ifla63/63gorm.htm. Reference given in Cathro 1999.) Of course, this solution would preserve only "static" elements of content: "dynamic" graphics, or dynamically generated content, could not simply be printed out.

[100] Hedstom 1995 also refers to some of the benefits digital storage can bring. This is a point made remarkably rarely in the literature! (See Hedstrom, M *Preserving Digital Information* in Fresko (ed) 1996.)

[101] Clarke 1997, paragraph 30 (op cit).

[102] A useful diagram representation of "refreshment" is given in Fresko, M *Results of the Comparative Study of Digital Preservation Guidelines*. See Fresko (ed) 1999 Digitisation of Library Materials. Report of

As one would expect, most national libraries are already making some use of this strategy. However, it is recognised that refreshment alone does not provide a long-term solution, because it does not address the problems that arise when the hardware and/or software required to run a publication become defunct. It therefore needs to be combined with *technology preservation*, or substituted with either emulation or migration.

3.2.4.3   Emulation

In emulation, the original hardware/software environment is imitated in successive hardware/software environments.[103] The attraction of this method of preservation lies at least in part in the prospect it seems to offer of maintaining the "look and feel" – in theory, at least, the complete functionality – of the resource.

Rothenberg's work has focused on this strategy.[104] Essentially, it involves encapsulating the data with the application software that created it and the supporting software required to run the application software, together with a description of the necessary hardware environment. This is used to run an emulator, a "virtual machine".

There is a heated debate about emulation at present, not least because of the expense it entails. The problems with the theory itself have been forcefully expounded by David Bearman: "… emulation is not a viable approach to preservation at this time and […] even Rothenberg does not suggest that it is. Electronic records that are not moved out of obsolete hardware and software environments are very likely to die with them".[105] He outlines what he regards as its functional inadequacies,[106] and also claims that emulation represents serious overkill as a preservation strategy for most electronic documents.

From our present standpoint, emulation seems to be a strategy for use in special situations, where the need to retain the "look and feel" of a resource is considered particularly important. There are significant risks involved in relying on emulation when proof-of-concept work is still required and when we cannot be certain that we will possess, in future, the technical skills and commercial underpinning to maintain such a strategy over the very long term.[107]

3.2.4.4   Migration

In this strategy, data is transferred wholesale from one hardware/software configuration to another, without attempting to imitate the original.[108] Migration is defined in the CPA/RLG report as "A set of organised tasks designed to achieve the periodic transfer of digital material from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation."

---

the Concertation Meeting and Workshop Held in Luxembourg, 14 December 1998. http://www.echo.lu/digicult/en/backgrd.html.

[103] Again, see the diagram in Fresko 1999 (p47).

[104] See for example Rothenberg 1995 (op cit) and Rothenberg 1996 (op cit).

[105] Bearman 1999, p1 (op cit).

[106] These are rooted in the fact that emulation does not adequately address "evidence", as he defines it (on which, see Bearman & Sochats 1995, op cit).

[107] See Hendley 1998, op cit (summarised in Feeney 1999, Chapter 4).

[108] A diagram is given in Fresko 1999, p46.

This strategy has many advocates,[109] and has the advantage of enabling the custodians of information to manage it within their own IT environment. This makes the implications of such a strategy all the more important to flesh out.

Data issues: Migration can only realistically be carried out using "raw" or "source" data – that is, data in the form it was in before being "compiled" or "built" into a publication. It will rarely, if ever, be possible to retrieve the raw data from offline media, since it is usually compressed, and sometimes encrypted,[110] which means one cannot simply extract it from the published item.

Where online publications are concerned, it is usually possible to download a site's front pages, and one might be able to download the data in certain formats (such as PDF or DOC or GIF) where that data is presented in a "passive" site.[111] However, active sites present greater problems, since the data does not exist in a format like HTML: it is held in a database and is generated "on the fly", or "dynamically", when the database is queried.

Suppose, for the sake of argument, that publishers deposited uncompiled data instead of, or as well as, the published item as it is made available commercially. Libraries would receive data in a mass of formats, many of them proprietary. We are far from a situation in which all publishers use application software-independent formats, such as SGML and XML,[112] for their electronic publications, which are widely regarded as the cornerstone of successful migration through changing technological regimes.[113]

Standards like SGML would certainly assist migration. Even if the standard itself does not last "forever" (which it will not), it should be possible to establish a standard migration path which is more likely to last permanently.

However, it is highly unlikely that, in the short term at least, there will be a large volume of SGML encoded data available from publishers for deposit.[114] Furthermore, we should note that SGML files are compiled in relation to a Document Type Definition (DTD). There are (potentially) as many DTDs as there are publications[115] and the DTD for each publication is

---

[109] See, for example, Bearman 1999 (op cit) and Beagrie & Greenstein 1998 (op cit).

[110] Some forms of "encryption" might even be added to documents without publishers' awareness. For example, in the online journals context, it seems that some suppliers use a particular pdf function which effectively "encrypts" the data without the publishers realising that they are doing this.

[111] See Section 4.2.2 above on "active" and "passive" sites.

[112] In practice, SGML is not genuinely "application independent": most DTDs are written for *specific* applications. We do not know the extent to which the same will be true of XML.

[113] Coleman, J and Willis, D *SGML as a framework for digital preservation and access* (Washington: Commission on Preservation and Access, 1997) say: "When viewed from the perspectives of the preservation community and the digital librarian, SGML appears to be the best choice for archival document storage. SGML is standardized, portable, flexible, modular, and supported by the market. [Its] vendor neutrality, extensibility … and ability to manage large information repositories make it an attractive choice for both archival and retrieval purposes."

[114] This might change if XML becomes the standard mark up language for online publication; however, from what we know, we believe it will be many years before a real critical mass of publications is available for deposit in structurally encoded form.

[115] In practice, many are likely to be very similar.

subject to change with each release of the data.[116] Migration, even of data managed in SGML or XML, will entail the management of a great range of DTDs.[117]

It must be emphasised that this approach assumes that librarians will know and understand the structure of the data of all the resources in their collections in a way that they do not currently; there is no reason for librarians to know the form in which publishers hold their data for their printed publications. Management of content at this level has very serious resource implications for libraries.

This approach also raises considerable data maintenance issues from a publisher's point of view. The deposit of "uncompiled" or "raw" data involves the deposit of "intermediates", or versions of publications which are "intermediary" in the product development cycle, rather than the final publication itself. The management of intermediate versions has historically been difficult for publishers, and the maintenance of an intermediate whose content is identical to the finished publication, and which therefore should be deposited, is a procedure that publishers would need to master.

"Look and feel" and functionality: Migrating data to a new technological environment will necessarily involve losing the "look and feel" of the previous environment. This raises the important question we referred to above: to what extent does anyone have the right to change the visual presentation of another person's work? Even assuming that we were able circumvent this problem – for example, if legislation were to give libraries the right to make such changes if preservation could not be achieved without them – we need to formulate a framework within which to operate. How important is it for us to preserve the "look and feel" of any particular product in the long-term, and what *exactly* do we mean by that?

For example, it might be considered desirable to preserve those elements of the functionality and presentation which can reasonably be deemed part of the content or "message" of the publication. In practice, the deposit libraries will undoubtedly be constricted by cost as much as anything else, in which case they will need to direct their resources towards preserving the elements of functionality and presentation *most* central to the "message". Clearly, there are a number of rather serious value judgements that need to be made in this scenario, and it will be necessary to consider who will be make these decisions and with what frame of reference.

Links: Hybrid and online publications contain links to other online resources. The preservation of these links is taxing, and again relates to the vexed issue of an electronic publication's *definition*: what are its limits? When one document links to another, and the second to a third, and the third to a fourth, what should be the limits to what is preserved?[118] Does it depend on the nature[119] of

---

[116] To what extent the DTD changes depends on the publication.

[117] There are some who have hoped that a small number of DTDs could be developed to cover all publications of a certain type which would be adopted by all publishers – thus, for example, all publishers of physics serials would use a single DTD. This seems unrealistic, certainly in the short term. There would be many political and managerial difficulties involved in operating such a system; equally, there is the technical reality that one DTD will not in fact be suitable for all types of publication.

[118] In the DNEP-IWI project, the KB downloaded 50 web documents and found that one either decided to preserve a small number of links, or one ended up preserving "half the internet". The National Library of Canada, in the context of the online publications it collects through voluntary deposit, initially decided to attempt to maintain all links in the same domain as the resource in question, but found that even that was unrealistic. They have narrowed down their definition to links which are "integral" to the document.

the link, and on whether the material is substantially intellectually impaired if its links are severed? And again, what exactly might such a definition mean, and who would decide?

These are questions which all digital libraries and archives are addressing, and about which there is little consensus on "best practice" as yet. Librarians who have carried out studies in this area have found that, currently, it is impossible to maintain the integrity of all the links contained in documents they collect.[120] The PANDORA project implemented a PURL Resolver Service to address the problem of broken links, but recognises that a more robust solution is required in the longer term.[121] The period of voluntary deposit in the UK will need to grapple with all these questions.

It appears that substantial editorial expertise will be required to answer many of these questions on a case-by-case basis – and that a rules-based approach may not be adequate.[122]

Before moving on – a note of realism. Societies and libraries have never been able to preserve everything that future generations might wish they had. This discussion is not intended to imply that electronic media will enable us to do so either. As we have said before, libraries who have been collecting electronic publications for some years realise that there are already publications to which they cannot provide access, and that the number of inaccessible publications is likely to grow.

Not the least of the difficulties that will be encountered, when strategies for preservation have been decided, will be the costs involved.[123]

3.2.4.5   Rights issues
Any method of preserving digital resources involve preservers' manipulating them to some extent. This may be simply copying them onto another medium (in the case of refreshment), encapsulating and emulating them (in the case of emulation) or, more radically, extracting their content and migrating their software, thus changing, to some extent at least, their presentation and functionality (in the case of migration). There are clearly issues relating to underlying rights involved at each stage.

---

[119] For example, it might be argued that links which are references should not be preserved, on the basis that references in a print publication are not necessarily preserved.

[120] For example, see above, n105.

[121] See www.nla.gov.au/pandora/summary.html in conjunction with http://purl.nla.gov.au

[122] This may imply, in turn, that publishers will need to make certain judgements about what they, and their authors, might expect to be preserved, and provide information to libraries about that. The cost of making all the decisions that are implied will be very high to both libraries and publishers alike.

[123] Compare Foot 1995 (op cit): "Lack of resources has always stood in the way of the successful implementation of a preservation policy or strategy and will certainly do so no less for electronic material." (Indeed, as she goes on to say, it will be worse, because electronic preservation requires an ongoing, not once-off, commitment, as mentioned above, in the introduction to Section 4.2.4.) We do not yet have much knowledge of the costs involved in digital preservation. Some attempts are made in the CPA / RLG report to address the management of costs and finances, but conclusions are necessarily tentative. Hendley 1998 (op cit) emphasises the difficulty of *isolating* those costs which pertain solely to the preservation of a digital resource.

It is beyond our remit to discuss legal issues in any depth, but it is important to flag them for discussion in the appropriate arenas. Currently, in CDPA (1988)[124], there is a specific exception to copyright under library privilege,[125] such that a library may copy "any item in its permanent collection" for the purpose of replacement or preservation. A copy can also be made for another library to replace a copy "lost, destroyed or damaged" in that other library. However, libraries are not empowered to copy items under this clause where it is "reasonably practical" to purchase a replacement copy of the item. It is likely that some legal development will be required to make clear exactly what rights the deposit libraries have with respect to preserving electronic publications.

There is another rights related issue that should not be overlooked. Publishers may not control (and may therefore be unable to grant) the rights that the libraries will expect to be able to exercise in order to preserve digital resources. This not only relates to certain authorial rights in the content, it may equally apply to software that the publisher has licensed for their products. This is clearly a non-trivial issue and is one that it may only be possible to resolve through legislation.

3.2.4.6   *Documentation and metadata*
We will not presume to recommend which preservation strategy, or strategies, the deposit libraries should pursue. Whatever the decisions taken, however, one thing is clear: resources will need to be uniquely identified and described fully if they are to be preserved successfully. As Beagrie and Greenstein put it, "No digital archivist can successfully preserve a data resource that is not fully documented".[126]

As Bearman says, all preservation models share a common research agenda: we must move forward in specifying the metadata that will be required for them.[127] Much work has been done on metadata for preservation, and many possible approaches have been identified. However, there is as yet little consensus in this area.

CEDARS is currently investigating metadata requirements for preservation, and expects to publish a paper on required elements at the end of the Summer of this year. Even in draft form, it is an extensive list. A project paper published in August 1998 highlighted the metadata specifications of the PANDORA and the Pittsburgh NHPRC project as being of particular interest, as well as the OAIS taxonomy of "Preservation Description Information".[128] In addition, it notes the work of the EBU-SMPTE Task Force, and the UPF and UBC Project. In addition, the

---

[124] Copyright Designs and Patents Act, 1988.

[125] See section 42.

[126] Beagrie and Greenstein, December 1998, p 28. Lievesley made a similar point in her address to the Warwick conference in 1995, *Strategies for Managing Electronic Archives*: "The preservation of documentation is also critical, as *data without its documentation is greatly reduced in worth if not entirely worthless*" [our emphasis]. (See Fresko 1996.)

[127] See Bearman 1999 (op cit)

[128] See Day 1998 (op cit). This includes metadata on provenance, content, context and fixity.

RLG Working Group on the Preservation Issues of Metadata has specified a set of 16 metadata elements for the preservation of digital images.[129]

In addition to these metadata specifications, preservers of digital information will need to consider the management of the metadata itself as it grows (for example, as metadata is created to log a resource's preservation history). This becomes all the more difficult when managing iterated instantiations of migrated resources over time.

3.2.5 Record Creation and Bibliographic Service Provision

There are others much more qualified than ourselves to discuss the creation of records and bibliographic listings within libraries. Current cataloguing rules may prove to be adaptable to offline publications, albeit with some deficiencies.[130] It is the cataloguing of online publications which librarians expect to cause more upheaval.

However, the ease and efficiency with which libraries will be able to create records and listings for deposited publications will depend critically on the way in which publishers make available the relevant information. It is this aspect of cataloguing which is a particular concern to us in this paper. It is an issue we addressed above in Section 3.2.3.2, relating to the initial accession of publications on receipt at the point of deposit. Record creation and bibliographic services can be seen as an extension of that initial registration process, and many of the issues discussed in relation to registration are equally pertinent here.

Given the quantity of data that we expect will, in time, be processed through the national archive, efficiency and productivity will be essential for the deposit libraries in these activities.[131] This was a point highlighted by all the national librarians we spoke to. It is critical, therefore, that metadata is as well-formed as possible at the point of creation.[132]

---

[129] See RLG Working Group on Preservation Issues of Metadata, May 1998 *Final Report* http://www.rlg.org/preserv/presmeta.html

[130] For example, AACR2 does not provide sufficiently for the categorisation of electronic products at present. (In the CD-ROM Demonstrator project CD-ROMs were simply identified as "computer files". Concerns about this were noted in Clarke 1997, paragraph 52.) Other countries use catalogue formats developed from MARC: Die Deutsche Bibliothek, for example, use the PICA-ILTIS format, which has been extended to cover electronic offline and online publications, and the Koninklijke Bibliotheek use the MARC-like format Pica Plus.

[131] Of course, the libraries will address definitions of "efficiency and productivity" internally. The CD-ROM Demonstrator project provides some relevant figures here: it was found that the initial registration of CD-ROMs took about the same amount of time as the registration of books (approximately 7 or 8 minutes for the first part of a serial publication and 4 minutes for a monograph), whereas the creation of a catalogue record took significantly longer (67 minutes for a CD-ROM, compared with 39 minutes for a book). See Clarke 1997, paragraphs 10 and 38. In addition, the comments of the National Digital Library Program on the feasibility of cataloguing at various levels of granularity are interesting (see Arms, April 1996, p6f).

[132] As emphasised above and, for example, in the CPA/ RLG report (see especially n13).

3.2.6   Reader Services: access to deposited publications

3.2.6.1   Access terms

The terms under which readers will be allowed to access deposited publications are being keenly debated in the UK.

The arrangements in place at national libraries in other countries can offer us some guidance, but it is striking that *many of these arrangements are in the process of being changed* on the basis of experience.

For example, the Library of Congress is currently seeking substantially to refine its "boilerplate" approach; publishers will be asked to agree that their material will be accessible to users according to one of three or four standard sets of terms and conditions.[133] Resources deposited by law are currently logged by the Library's Copyright Office and then distributed to the relevant collection (usually a specialist reading room), and from this point onwards they are treated just like other resources in that collection. Usually, therefore, resources are confined to and accessed from workstations in the relevant area of the Library.[134]

The Koninklijke Bibliotheek currently allows networked access to (voluntarily) deposited publications "on site", but it anticipates that "on site" might in the future be defined as "being a registered user" of the library. At the Bibliothèque nationale de France, the deposit collections are not networked, but are installed on individual workstations on request. Restricted conditions of access are logged in the bibliographic record and these are checked at the installation.

The code of practice for the voluntary period of deposit in the UK is likely to suggest that publishers select an access "option" (one option is likely to be the default). Publishers might allow tightly controlled network access between the deposit libraries if they wish, or they might prefer to allow minimum access, which is defined as access, via a secure, closed network, to a single concurrent user physically present in a single deposit library.

It will be one necessary function of the voluntary deposit period to define much more rigorously the precise forms that this limited networking will take.[135]

3.2.6.2   Metadata relating to terms of availability

It follows that, in the absence of a single overarching agreement covering access to all deposited publications, a crucial subset of the metadata deposited by publishers will relate to the terms under which deposit is being made.

The development of ways of describing rights agreements (which is essentially what "terms of availability" are) is still in its infancy for the publishing industry.[136] Although the necessary logical structures for defining rights agreements will be one of the core deliverables of the <indecs> logical model, this aspect of the model is still under development in September 1999. Furthermore, the development of a logical way of expressing agreements does not imply that,

---

[133] This will cover both resources purchased by the library and those obtained under the US mandatory deposit law.

[134] There is a notion that the Library of Congress dedicates a single machine to deposited resources, but this is not quite the case.

[135] Eighth draft, p2.

[136] And, indeed, for the "content industries" in general.

suddenly, there will be massive amounts of data available to express them. Publishers currently have little or no experience of the management of rights databases of this kind.

Ultimately, there may be a need to develop very sophisticated models for the management of terms of availability for deposited content. In the short term, there can be no expectation that this will be possible. It will be necessary for a small set of "terms of availability" attribute values to be developed as part of the work in developing the deposit metadata set.

3.2.6.3   *Usage*

One aspect of deposit which is not our focus, but which is worth mentioning, relates to managing the use made of collections. The deposit libraries will confront many issues here, including the extent to which they wish to monitor the use made of deposit collections. It is worth bearing in mind the costs involved in doing so.[137] Another issue will be the level of reader services required. In this connection, the CD-ROM Demonstrator project noted that more documentation and explanation was required for users than predicted, and that significant technical support for staff is needed for electronic publications.[138]

4   Conclusions

The concept of centralised deposit of publications was first introduced in response to the introduction of a new technology: the printing press.[139] The operation of deposit schemes is, at present, linked at a profound level to that technology. We are now experiencing the early phases of another revolution in technology which will have at least as great an impact as printing did five centuries ago.

We are beginning to understand what it means to publish electronically. It is entirely appropriate that we should completely re-assess the process of legal deposit in the light of that new technology; it is not unreasonable to expect that attitudes, methods and practices will need to change. This paper has attempted to guide the reader through some of the issues which we believe need to be considered afresh in the light of new publishing technologies.

In considering technical standards relating to deposit of electronic publications, we have seen the extent to which technology issues are inextricably bound up with issues of policy and strategy. It is essential that the discussions which take place relating to policy are informed by a clear understanding of the technology issues and *vice versa*.

It is clear from our discussions with publishers that, in general terms, they would be willing to co-operate with requests from the copyright libraries, although a number of them mentioned the relatively high compliance costs which would be incurred, particularly for cumulative online publications.[140] Those we spoke to in publishing houses were, for the most part, in technical roles

---

[137] Those librarians we spoke to whose institutions do monitor usage give a picture of slowly growing demand for publications.

[138] See Clarke 1997, paragraphs 60ff.

[139] A very useful overview of the history of legal deposit can be found in Ratcliffe, F W *Legal deposit: Not a copyright issue – a cultural legacy for the future* Logos 2/2 1991.

[140] One publisher, who is planning a frequently updated cumulative online scientific resource suggested to us that preparing material for deposit could require the dedication of half-a-man-day every two weeks – a not insignificant commitment.

and would not be able to incur these costs without higher management approval. We did not investigate the policy issues that this implies since this was outside our brief. However, we can expect that measurement of the cost of compliance will be a significant issue for publishers during the voluntary deposit phase – as, no doubt, will measurement of the cost to the deposit libraries.[141]

The ability of publishers to co-operate technically is hampered by a lack of clarity about what it would actually mean to deposit electronic publications. We have in the text identified a number of primary questions which need to be addressed:

- How will the deposit libraries define and articulate the purpose of deposit of electronic publications (see Section 3.1)?

- What strategy or strategies for preservation will they adopt (3.2.4)?

- How will the libraries identify publications eligible for deposit – or will this be the sole responsibility of publishers (3.2.1)? And how will publishers themselves be defined in the online environment (3.2.3.2)?

- To what extent will it be acceptable to modify resources if preservation cannot be achieved without modification (see 3.1 and the technical discussion in 3.2.4)?

- Who will be empowered to make what are effectively editorial decisions relating to deposited publications, and within what frame of reference (3.2.4.4)?

- How will items for deposit be delivered from publishers to the libraries, particularly online publications (3.2.2)?

- What should be delivered – intermediates or the final, published product (3.2.4.4) and how will this be decided?

As explained in 3.2.3, a unique product identifier is not yet in "standard" use, and other information which it is critical for the libraries to know about a resource is not available in a standardised fashion. Discussions are required to decide *what* metadata is required at the point of accession, and both the format and the protocols by which it might be delivered. To these deliberations, the work of the <indecs> project, the International DOI Foundation, CEDARS and BIBLINK will be significant contributors. We recommend that the deposit libraries begin work with publishers as soon as possible in developing appropriate metadata schemas for deposited content. Further, once access standards have been agreed, it will be essential for standard means of expressing these terms of availability in metadata to be agreed.

It is clear that this combination of technical and policy discussions will not be answered in the very short term. However, we believe that it should be the aim of the voluntary deposit scheme actively to seek robust answers to these questions in advance of any move towards the establishment of a legal deposit scheme. To this end, we recommend the establishment of a joint committee, made up of technical representatives of the deposit libraries and a broad cross section of the electronic publishing community to attempt to find mutually acceptable solutions to the questions posed. This committee should be established under the auspices of Book Industry

---

[141] We make no attempt in this document to address the question of how the costs associated with the preservation of electronic publications will be met, while recognising that, whatever strategy or strategies for preservation are finally adopted by the deposit libraries, these costs will not be negligible.

Communication which is uniquely placed to ensure that the interests of both communities are properly served.

It is our hope that this report will form the basis for their initial agenda and will provide a starting point for their deliberations.

References

Alkhoven, P *Metadata Needed For Handling Digital Content* In Fresko (ed) 1999

Arms, C R *Historical Collections for the National Digital Library. Lessons and Challenges at the Library of Congress* D-Lib Magazine April and May 1996.
http://www.dlib.org/dlib/april96/loc/04c-arms.html and
http://www.dlib.org/dlib/may96/loc05c-arms.html

Beagrie, N and Greenstein, D *A Strategic Policy Framework for Creating and Preserving Digital Collections* British Library Research and Innovation Report 107. 1997
The final draft, dated 14 July 1998, is available at http://ahds.ac.uk/public/srg.html

Beagrie, N and Greenstein, D *Managing Digital Collections: AHDS Policies, Standards and Practices* December 1998, Consultation Draft.
http://ahds.ac.uk/public/srg.html

Bearman, D and Sochats, K. *Metadata Requirements for Evidence* 1995
http://www.lis.pitt.edu/nhprc/BACartic.html

Bearman, D *Reality and Chimeras in the Preservation of Electronic Records* D-Lib Magazine April 1999.
http://www.dlib.org/dlib/april99/bearman/04bearman.html

Bennett, J C *A Framework of Data Types and Formats, and Issues Affecting the Long-Term Preservation of Digital Material* British Library Research and Innovation Report 50. 1997

Boyce, P B *Costs, Archiving and the Publishing Process in Electronic STM Journals* Against the Grain November 1997

The British Library *Proposal For The Legal Deposit Of Non-Print Publications To The Department Of National Heritage From The British Library* 1996

Cathro, W S *Digital Libraries: a National Library Perspective* January 1999

Clarke, A *British Library Legal Deposit CD-ROM Demonstrator Project* May 1997

Coleman, J and Willis, D *SGML as a framework for digital preservation and access* Washington: Commission on Preservation and Access, 1997

Conway, Paul *Preservation in the Digital World* 1996 www.clir.org/pubs/reports/conway2

The Data Archive, University of Essex *An Investigation Into The Digital Preservation Needs of Universities And Research Funders: The Future Of Unpublished Research Material* British Library Research and Innovation Report 109. 1998

Day, M, June 1998 *CEDARS: Digital Preservation and Metadata*
http://www.ukoln.ac.uk/metadata/presentations/delos6/cedars.html

Day, M, August 1998 *Metadata for Preservation. CEDARS Project Document AIW01*
http://www.ukoln.ac.uk/metadata/cedars/AIW01.html

Department of National Heritage *Legal Deposit Of Publications: A Consultation Paper* 1997

Feeney, M *Digital Culture: Maximising The Nation's Investment. A Synthesis of the JISC/NPO Studies on the Preservation of Electronic Materials.* The National Preservation Office: 1999

Foot, M *A Preservation Policy for Digital Material: A Librarian's Point of View* In Fresko (ed) 1996

Fresko, M (ed), *Long Term Preservation of Electronic Materials: A JISC/British Library Workshop as part of the Electronic Libraries Programme (eLib)* 27th and 28th November 1995 at the University of Warwick. British Library R&D Report, 6238. 1996.

Fresko, M (ed), January 1999 *Digitisation of Library Materials. Report of the Concertation Meeting and Workshop Held in Luxembourg, 14 December 1998.* http://www.echo.lu/digicult/en/backgrd.html

Fresko, M *Results of the Comparative Study of Digital Preservation Guidelines* In Fresko (ed) 1999

Fullerton, J *Developing national collections of electronic publications: issues to be considered and recommendations for future collaborative actions* 1998 http://www.nla.gov.au/nla/staffpaper/int_issu.html

Graham, P *Preserving the Digital Library* In Fresko (ed) 1996

Green, B & Bide, M *Unique Identifiers – a brief introduction* (revised version March 1997) London: Book Industry Communication. http://www.bic.org.uk/bic/uniquid.html

Haynes, D, Streatfield, D, Jowett, T and Blake, M *Responsibility For Digital Archiving And Long-Term Access To Digital Data* British Library Research and Innovation Report 67. 1997

Hedstrom, M *Preserving Digital Information* In Fresko (ed) 1996

Hendley, T *Practical Implications: Long-Term Preservation of Electronic Materials* In Fresko (ed) 1996

Hendley, T *Comparison of Methods and Costs of Digital Preservation* British Library Research and Innovation Report 106. 1998

IFLA Study Group *Functional Requirements for Bibliographic Records* K G Saur: 1998

Kahle, B, *Archiving the Internet* Scientific American, March 1997 http://www.archive.org/sciam_article.html

Keefer, A *Preservation of Electronic Publications* Presentation to the SLA Mediterranean Conference at Barcelona, 26-27 February 1999

Leggate, P *Internet Library of Early Journals Project* In Fresko (ed) 1999

Library and Information Commission *New Library: The People's Network* 1997

Lievesley, D *Strategies for Managing Electronic Archives* In Fresko (ed) 1996

Matthews, G *Preservation Management* LIBS: Library and Information Briefings, 73, August 1997

Matthews, G, Poulter, A and Blagg, *E Preservation of Digital Materials Policy and Strategy Issues for the UK: Report of a Meeting on the CPA/RLG Report, December 1996* British Library Research and Innovation Report 41. 1997

National Library of Australia *Voluntary Deposit Scheme for Physical Format Electronic Publications*
http://www.nla.gov.au/policy/vdelec.html

Paskin N *Information Identifiers* Learned Publishing 1997: Vol 10 No 2
http://www.elsevier.nl/locate/infoident

Paskin N *DOI: Current status and outlook* D-Lib 1999: Vol 5 No 5
http://www.dlib.org/dlib/may99/05paskin.html

RLG Working Group on Preservation Issues of Metadata, May 1998 *Final Report*
http://www.rlg.org/preserv/presmeta.html

Ratcliffe, F W *Legal deposit: Not a copyright issue – a cultural legacy for the future* Logos 2/2. 1991

Ross, S and Gow, A *Digital Archaeology: The Recovery of Digital Materials At Risk* British Library Research and Innovation Report 108. 1999

Rothenberg, J. *Ensuring the Longevity of Digital Documents* Scientific American Vol. 272, Number 1. January 1995

Rothenberg, J., *Metadata to Support Data Quality and Longevity*. Proceedings of the 1st IEEE Metadata Conference, 16-18 April 1996.
http://www.computer.org/conferen/meta96/rothenberg_paper/ieee.data-quality.html

Waters, D and Garrett, J *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group*. Commission on Preservation and Access, 1996.
http://www.rlg.org/ArchTF

Zillhardt, S *Bibliotheca Universalis Assessment* In Fresko (ed) 1999

List of acronyms used in report

AHDS	Arts and Humanities Data Service
BnF	Bibliothèque nationale de France
BL	The British Library
CD-ROM	Compact Disc - Read Only Memory
CEDARS	CURL Exemplars In Digital Archives
CPA	Commission on Preservation and Access
CURL	Consortium of University Research Libraries
DAWG	The Digital Archiving Working Group
DC	Dublin Core
DDB	Die Deutsche Bibliothek
DOI	Digital Object Identifier
DTD	Document Type Definition
EBU	European Broadcasting Union
eLib	Electronic Libraries Programme
HTML	Hyper-text Mark-up Language
IETF	Internet Engineering Task Force
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
ISO	International Organisation for Standardisation
JISC	Joint Information Systems Committee
KB	Koninklijke Bibliotheek, The Netherlands
LC	Library of Congress, USA
MOAII	Making of America, Second White Paper
NASA	National Aeronautics and Space Agency
NBA	National Bibliographic Agency
NDLP	National Digital Library Program (at the Library of Congress)
NDLF	National Digital Library Federation
NEDLIB	Networked European Deposit Library
NHPRC	National Historical Publications and Records Commission
NLA	National Library of Australia
NPO	National Preservation Office
OAIS	Open Archival Information System

OCLC	Online Computer Library Center
PANDORA	Preserving and Accessing Networked DOcumentary Resources of Australia
PURL	Persistent Uniform Resource Locator
RDF	Resource Description Framework
RLG	Research Libraries Group
SGML	Standard Generalised Mark-up Language
SMPTE	Society of Motion Picture and Television Engineers
SSHRCC	Social Sciences and Humanities Research Council of Canada
TEI	Text-Encoding Initiative
UBC	University of British Columbia
ULCC	University of London Computing Centre
UPF	Universal Preservation Format
URL	Uniform Resource Locator
URN	Uniform Resource Name
W3C	World Wide Web Consortium
XML	Extensible Mark-up Language

Acknowledgements

We would like to thank the following people for their valuable contributions to this study, while acknowledging that any errors or misinterpretation are of course our own.

John Akeroyd	South Bank University
Michael Alexander	British Library
Caroline Arms	Library of Congress
Christine Bossmeyer	Die Deutsche Bibliothek
Nancy Brodie	National Library of Canada
Gesche Buecker	Waterlow Specialist Information Publishing Ltd
Jasmine Cameron	National Library of Australia
Reg Carr	Bodleian Library, Oxford
Robina Clayphan	British Library
Adrian Cunningham	National Library of Australia
Melissa Dadant	Library of Congress
Anne Dixon	Institute of Physics Publishing
Jacques Faule	Bibliothèque nationale de France
Chris Fell	Cambridge University Press
Jack Flavell	Bodleian Library, Oxford
Carl Fleischhauer	Library of Congress
Lloyd Fletcher	Institute of Physics Publishing
Wendy Frankland	British Library
Peter Fox	University Library, Cambridge
Gina Fullerlove	Macmillan Reference Ltd
Andrew Green	National Library of Wales
Brian Green	Book Industry Communication
Dan Greenstein	AHDS
Dr Rhidian Griffiths	National Library of Wales
Paul Guy	British Library
Tony Hammond	Academic Press, Harcourt Science and Technology Company
Helen Henderson	Dawson Publisher Relations
Volker Henze	Die Deutsche Bibliothek
Doug Hodges	National Library of Canada
David Inglis	British Library
Hans Jansen	Koninklijke Bibliotheek
Hans Liegmann	Die Deutsche Bibliothek
David Lightning	Lightning Software Development
Karen Little	Chadwyck-Healey Ltd
Fiona Macdonald	CRC Press
Gavin McDonald	Blackwell Publishers Ltd
Ian McGowan	National Library of Scotland
Cliff Morgan	John Wiley & Sons, Ltd
Keith Nettle	Publishers Association
Trudi Noordermeer	Koninklijke Bibliotheek
Rene Olivieri	Blackwell Publishers Ltd
Mary Grace Palumbo	Dawson Information Services Group

Norman Paskin	International DOI Foundation
John Peacock	Macmillan Press
Andrew Pearson	Vignette
Stephen Pocock	Chadwyck-Healey Ltd
Rhona Richard	Blackwell Publishers Ltd
Chris Rusbridge	JISC
Kelly Russell	CEDARS
Ute Schwens	Die Deutsche Bibliothek
Bill Simpson	Library of Trinity College, Dublin
Tom Smail	Agent, Copyright Libraries Agency
Geoff Smith	The British Library
Neil Smith	British Library
Sian Smith	Macmillan Publishers Ltd
Johan Steenbakkers	Koninklijke Bibliotheek
Andy Stone	CEDARS
John Strange	Blackwell Science
Colin Webb	National Library of Canada
Dave Wilkie	British Library
Titia van der Werf	Koninklijke Bibliotheek
Sonia Zillhardt	Bibliothèque nationale de France