STANDARDS FOR ELECTRONIC PUBLISHING

An overview

A report for the NEDLIB Project
by

Mark Bide & Associates
mark@markbide.co.uk

August 2000

Contents

Executive Summary

1. This report was commissioned by the NEDLIB project as a supplement to its *Process Model for a Deposit System for Electronic Publications.*

2. Our intention is assist librarians to a better understanding of some of the issues that they will face in managing the deposit and preservation of electronic publications. The report is based on interviews with 14 European publishers from different publishing sectors.

3. The publishing industry – even the book and journal publishing industry – is a heterogeneous community. Processes in different sectors – consumer, educational, professional, academic – are often significantly different one from another. This can make generalising about trends difficult.

4. Production and production processes do not lie at the heart of publishing's main concerns. Most publishing production has traditionally been outsourced.

5. Publishers adopt standards only where there are good commercial reasons for doing so. They will not adopt standards solely to aid the process of deposit and long-term preservation (unless there are good commercial reasons for them to do so).

6. Interchange of metadata between publishers and libraries will be an essential element in managing the deposit of electronic publications. Publishers are becoming increasingly concerned about metadata and metadata standards, driven by the commercial imperatives of selling their products – both physical and digital – over the Internet. New standards are being rapidly adopted and may provide a basis for future sharing of metadata. However, publishers have much to learn in this respect.

7. Despite the success of some existing CDROM subscription publications, particularly in professional markets, CDROM and other offline publishing is in decline, with publishers favouring online distribution. This is good news for deposit libraries, insofar as CDROM publishing has been much less standardised than is the case with online (and the content of CDROMs is often encrypted).

8. Online, publishers are using standard Web browsers as their primary interface, although frequently users require additional applications (of which the PDF reader plug in is the most common). In online academic journals, there is a considerable growth in the publication of "supplementary content" supplied by authors. These are published in whatever file format the author submits in – often Excel or Access, but this is uncontrolled and only standardised to the extent that authors use "standard" applications.

9. A surprisingly high proportion of the publishers in our sample are now holding significant amounts of text SGML/XML archives.

10. Almost all of the publishers in our sample are delivering page images to users; PDF is near-universal as the format of choice for delivering page images.

11. There is little consistency between publishers in their use of formats for multimedia elements (although these remain relatively little used).

12. Many published products are dependent on middleware for their presentation to users (providing on the fly SGML to HTML conversion, for example); this middleware may often be part of the hosting service, and not under the direct control of the publisher. Educational content may be particularly dependent on integrated learning systems.

13. Publishers are showing considerable interest in ebooks, although the technology remains immature and standards are somewhat fluid. There remains a fundamental divide between those applications that depend on pre-formatted pages (PDF) and those that use the Open eBook, XML-based approach. The latter is appropriate for simple texts, but cannot manage complexity of content. ebooks will create particular problems for deposit libraries both because of diversity of content formats and because of the use of encryption and other security devices.

14. Although some very large publishers are managing more of their technical production and content management processes in house, we believe that it is likely that most will continue to look to outsourcing of these functions.

15. The explosion of publishing that is the public World Wide Web raises some particular problems for deposit libraries, not least in developing selection mechanisms.

16. We see reasons for deposit libraries to be optimistic about the adoption of standards in publishing, not least in the increasingly widespread use of SGML. However, this optimism needs to be tempered with realism – the fact that publishers are using more SGML does not solve many of the difficult questions relating to deposit and long-term preservation.

17. Both online and ebook publishing will see the “network effect” take greater and greater hold on standards. The market will enforce standardisation wherever lack of standardisation creates inefficiency for the end user. Inefficiency for the end user and inefficiency for the deposit library are not necessarily entirely congruent. However, the trends are encouraging.

1 Introduction

NEDLIB is an EC-funded project, involving as partners the National Libraries of several countries of the European Union.[1] The project is due to complete its work in December 2000. Its primary output is a *Process Model for a Deposit System for Electronic Publications*; this is accompanied by a set of *Deposit Guidelines* and a proposal relating to *Metadata for Long-Term Preservation.*

This report, **Standards for Electronic Publishing**, was commissioned by NEDLIB in June 2000, as a supplement to the *Process Model.* Its aim is to provide an overview of the extent to which publishers are using common standards in their electronic publications (both in the publications themselves and in the processes which lead to publication). It is anticipated that a review of this topic will assist librarians to a better understanding of some of the issues that they will face in managing the deposit and preservation of electronic publications.

The report was developed through a series of structured interviews (conducted around an emailed questionnaire) with 14 European publishers, supplemented by information focussed specifically on ebooks from US-based contacts.[2] (Although a number of our respondents in Europe are becoming involved in ebook production, the majority of activity and expertise today is in the United States).

Deliberately, most of the contacts that we made were with larger (book and journal) publishers, since these organisations are more likely to employ specialists with the necessary experience and expertise to answer the questions that we wished to put to them.

Prima facie, it may also appear that the output of larger publishers is more significant to the deposit libraries since a fairly high proportion of the published output – particularly of academic books and journals – is concentrated in relatively few publishing houses. However, technology is facilitating a move in the direction of greater fragmentation of publishing; in the next decade, smaller publishers may come to assume rather greater significance in terms of the complexity of managing deposit of electronic publications.

In view of the relatively small sample size, the report is not intended to provide statistically valid *quantitative* assessments of the extent to which standards are being adopted by publishers, but we have aggregated the data that we have collected to the extent that this seems useful.

We have included at the end of the report (Chapter 7) a brief note on the content of the public Web, based on harvesting data from the Netherlands and Sweden.

[1] See http://www.kb.nl/coop/nedlib/ for further details.

[2] Our sources are acknowledged in Chapter 9.

2 Terms and definitions

2.1 Definitions

We do not believe that it is necessary (or useful) in a report of this kind to provide a glossary or definitions of commonly used technical terms and acronyms like SGML or PDF. It is assumed that all of our readers will be sufficiently familiar with the technology to understand such technical terms as we will unavoidably use.

However, some initial definition of certain terms is necessary to avoid misunderstanding. These are terms where there may be some ambiguity of meaning, because of the lack of precision with which these terms are sometimes used.

We will talk about the following categories of publication:

- **offline** electronic publications, by which 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.[3] (Note that this definition specifically does not include ebooks.)

- **hybrid** electronic publications, by which we mean offline publications which contain links to online material.

- **online** electronic publications, by which we mean published resources accessible on the Internet or on proprietary networks.

Online publications can be further characterised by the “fixity” of their publication. We use the following terms to describe fixity:

- **static resources**, those whose form and/or content is recognised as substantially *fixed* throughout at the point of publication and throughout its lifecycle.

- **cumulative resources,** those whose content is being *added to* throughout its lifecycle (but where additional content elements are themselves substantially fixed); these are probably the most common online resources that are of interest in this report, since they are broadly the equivalent of serial publications.[4]

- **dynamic resources**, those whose form and/or content change continuously or ‘dynamically’ throughout its lifecycle. Dynamic resources will be addressed very little in this report, since the issues relating to their deposit are not primarily standards related.

We excluded **ebooks** from our definition of offline publications deliberately; they do not fit easily into any category, since they are typically distributed online but consumed offline.[5]

Current publishing-community efforts to define ebooks are proving complex. The latest working definition (V2.0) of an ebook from the AAP[6] ebook initiative is:

[3] They are variously described in the literature as “offline”, “hand-held”, “portable” and “packaged” electronic publications.

[4] This definition is conceptually similar to the AACR definition of a **continuing resource**: A bibliographic resource that is issued over time, usually with no predetermined conclusion. Continuing resources include serials and integrating resources.

[5] This is an interesting distribution model that has been developing in a number of different media, most notably perhaps in the widespread adoption of the MP3 format for the (legitimate and illegitimate) distribution of music. It has been argued that publishers might favour such a distribution model, but it has some distinct disadvantages from the publisher’s standpoint, not least in that commercial exploitation requires much more attention to the security of the content (see Section **Error! Reference source not found.**). Rather it appears to us that it accurately reflects the way in which users consume (and expect to consumer) many different types of content, without the necessity of being online. This distinction between “online” and “offline” may break down, if and when permanent and continuous wireless connection to the Internet (or perhaps its successor networks) becomes the norm. However, this will not happen in the immediate future and the “distributed ebook” is a phenomenon that will be with us for some time to come.

> *"a Literary Work in the form of a Digital Object consisting of one or more standard Unique Identifiers and [a] Monographic body of content, intended to be published and accessed electronically."*[7]

This definition is further qualified by definition of most the terms used. Most importantly, "literary work" is defined by reference to copyright legislation and "monographic" is used to distinguish an ebook from an e-serial (rather than in the narrower, traditional sense of a printed work on a single, usually learned, topic).

It is this definition that we will use in our report.[8]

The difficulty of naming and definition in this particular area of electronic publication should serve to alert us to the generic problem of unambiguous communication between libraries and publishers (and other interested parties) in the electronic publishing arena. Clear definition of terms is essential.

The final definition we will introduce at this point is **metadata**. We use the term in the broad sense of "descriptive data", encompassing what we commonly understand to be "bibliographic information" and also the more complex technical data that is essential to the management and preservation of electronic publications.

2.2 Presentation or content?

For as long as we can remember, there has been a fundamental division in electronic publishing between an approach based on logical content mark-up and an approach based on form (or presentation). In print-on-paper publishing, these two approaches are so intimately bound together (in typographic conventions and the traditional layout of front and back matter of books, for example) that we rarely think about them (unless, as sometimes happens, an inappropriate typographic treatment imposes itself between the reader and the meaning of the text).

However, typographic conventions prove to be unexpectedly opaque to computer interpretation. The more sophisticated manipulation of text that is possible if it is marked up for logical content (SGML and its derivatives) is much more difficult to achieve using presentation-based electronic formats (where PDF has become dominant – see section 3.3).

Nevertheless, following the PDF route has some substantial advantages to publishers. PDF is extremely inexpensive and easy to produce as a by-product of the traditional print-on-paper publication process (and indeed increasingly is itself becoming an integral part of print-on-paper workflows). PDF also provides publishers with a suitable format for managing content for print-on-demand applications, which are of growing commercial significance.

Producing text that is marked up for form has traditionally been significantly less expensive than producing the same text marked up for content (and then formatting it for print). Although the additional cost of producing "neutrally coded" text is now much less than it was, the number of suppliers able to produce SGML of a sufficiently high quality for reliable and consistent publication remains restricted (for complex text, at least).

On the other hand, there is self-evidently something of a mismatch between online publication and the use of the "printed page" presentation metaphor. PDF pages often present

[6] Association of American Publishers.

[7] We have some concerns about this definition of an ebook since it seems to conflate two related but different entities – the literary work (an abstraction) with the digital file (a manifestation of that abstraction). However, that concern need not detain us here, since it is not immediately relevant to our present discussion. The AAP definition also includes a secondary meaning that relates to ebook hardware devices that similarly need not concern us.

[8] There has been some discussion of the use of the term ebook to encompass electronic files that are intended for electronic dissemination but consumption by print-on-demand. Since books that are printed are likely – for the time being at least – to be deposited in libraries in printed form, we have not seen it as necessary to concern ourselves with this issue for the purposes of this report. However, the whole question of books published solely on demand or in ebook form is likely to become a significant issue during the forthcoming decade. See Chapter 4.

information rather badly on the standard computer screen and the page metaphor is arguably entirely unnecessary in an electronic environment.

However, by publishing in PDF, publishers have been able to maintain complete control over the format of their electronic publications (and through this of their branding). They can also ensure complete conformance in content between print and electronic versions (at least while parallel publication continues, although increasingly electronic versions of publications include content that the printed version does not).

There is another advantage to publishing in PDF that may reduce over time (but is certainly significant today). PDF allows the publisher easily to represent arbitrarily complex text (mathematics, chemistry and tabular material) in a way that standard methods for rendering marked up text cannot easily match. Although there are developments (like MathML) that may ultimately overcome these problems, publishers that have followed the “full text SGML” route to online publication currently have little option but to present complex text either as graphics or by using proprietary technology.

This question of “presentation or content” remains essentially unresolved in much electronic publishing. It has reappeared once again in the ebook domain (see Chapter 6). It seems unlikely to be resolved in the near future, and publishers are likely to have to maintain multiple parallel formats for the foreseeable future.[9]

[9] The *British Medical Journal*, for example, is published each week in both HTML and PDF versions – as well as in print.

3 Publishing – some background issues

3.1 A heterogeneous community

It is almost always misleading to talk about "publishing" as if it were a single industry; there are many different publishing industries, loosely grouped together by their use of a common medium for dissemination – print on paper. "Publishing" is conventionally (and relatively uncontentiously) broken down into major sub-categories – newspapers, magazines, books, journals. Even at this high level of granularity, some significant problems begin to appear at the margins – in defining (for example) when a magazine becomes a journal.

Our major focus in this report is "book and journal" publishing – but this categorisation masks considerable differences between different sectors. There are at least four major categories of book publishing: consumer (trade); educational; professional; and academic. Each of these has very different business models and very different businesses.

As we move further into the network environment, various aspects of our approach to categorising publishers must start to change. There is considerable convergence between different types of content (text, graphics music, audiovisual) as they come together for the first time in a common distribution environment. However, the differences between the different markets to which each of these types of publisher disseminates are in some ways further highlighted by the different ways (and speeds) with which different communities of users are adopting electronic dissemination.

There are, for example, some interesting sectoral differences already beginning to emerge in attitudes towards ebooks (on top of the strategic differences you might expect between different companies in the same publishing sector).

The publishers to whom we spoke for this report are primarily in professional and academic publishing; however, we included some publishers who publish educational (pedagogically structured) content and some who publish for general consumers (although none for whom this is their sole business). We will draw out such differences as we observe between these groups, although we must again stress that our sample size in insufficient for drawing statistically significant conclusions.

3.2 Publishing and the production process

We often observe a tendency (from outside publishing at least) to equate "the publishing industry" with the process by which an author's raw material is converted into a consumable product; this is the aspect of publishing that we mean when we talk about the "production process".

Although the origins of publishing and printing may have been very closely linked, for the greater part of the twentieth century at least, the majority of publishers took very little interest in their production process. Production was (and for many remains) a relatively unimportant part of their activity; most or all of it has been subcontracted to external suppliers, and the technical aspects of the process are of little interest to publishing management.

In recent years, those publishers who have felt the greatest impact of the digital revolution – particularly those in professional and STM publishing – have greatly increased their technological expertise. The type of larger publisher to whom we spoke for this report now employ people who can speak with some degree of confidence about SGML, for example. This would not have been uniformly the case five or ten years ago.

Nevertheless, publishers themselves emphatically do not equate publishing and production.

What is more, we have spoken to a somewhat non-representative sample of the publishing industry as a whole. Most publishers do not have the kind of expertise that exists among our target group, and have no desire to acquire it. They will publish in digital formats as the market demands that they do so; but (as they have for decades) they will continue to depend entirely on outsourcing for all the technological aspects of the production process.[10]

The significance of this, from the point of view of libraries with an interest in digital deposit and preservation, is that the people with whom they will find themselves dealing (once they move beyond the relatively small group of large publishers with whom they currently have contact) will not have much (if any) understanding of the technology involved. They will not manage their own content, but will depend on third parties. In an era of increasing globalisation, those third parties are as likely as not to be elsewhere in the world.

This is likely greatly to complicate dealings relating to the management of digital deposit.

3.3 Publishers' motivation for adopting standards

We will introduce immediately here one theme that will be – implicitly if not explicitly – recurrent throughout our report. Publishers' motives in adopting standards (or in not doing so) are the same as those that drive all of their decision-making; they are strictly commercial. The decision on whether or not to adopt a standard will be driven entirely (or almost entirely) by perceptions of their own commercial advantage.

In general, publishers do not produce their publications with an eye to their long-term preservation. They produce publications (in any medium) to sell them at a profit; long-term preservation may become significant to them, but only when the failure to accommodate its requirements in some way looks likely to have a significant impact on their commercial interest.

As one publisher revealingly said to us:

> *"I am not particularly optimistic [about the ability to access electronic publications] if you take the long term (>30 years?) view. I don't think publishers are well equipped or even well motivated to manage the long-term availability of their material – any material is only going to be cared for if it's financially viable. Once we can't get any money for something, we're not likely to worry about it. We should deliver whatever we can to the preservation libraries, but then it's over to them."*

Several of our other respondents explicitly stated that they considered long-term preservation to be "someone else's problem".

In the recent history of print-on-paper publishing, we can observe the direct analogy of the adoption of "acid-free" papers. When publishers of scholarly monographs were first asked (by libraries) to produce books on acid-free paper, the response from larger academic publishers was generally slow and uninterested. Acid-free paper tended to be more expensive and difficult to source.

This remained the case until competition in paper manufacture eliminated the price premium from acid-free stocks. At that point, when it became entirely painless to make the transition, publishers were happy to adopt acid-free paper and to earn whatever public relations value there was from being able to advertise the fact that they had done so. There was no commercial disbenefit and possibly (in terms of marketing, at least) a small commercial benefit. The decision became an easy one to make.

Commercial motivation is now driving publishers of one specific type of content – scholarly journals – to think harder about preservation issues. Major publishers (including, for example, Elsevier Science)[11] have made a commitment to moving their primary publication platform from print-on-paper to electronic publication in the very near future.

[10] There is anecdotal evidence that this does not apply to "new publishers" (such as one of our smaller respondents) that have become established primarily to publish electronically and whose business model and approach is therefore not hide-bound by a traditional print bias. Getting content suitable for preservation from these new publishers may prove more successful than from traditional publishers.

[11] The Managing Director of Science Direct, Elsevier Science's web-based content delivery service, said at a conference in June 2000 that their expectation is that **all** of their subscribers will migrate to electronic delivery within 2 or 3 years.

It has been widely posited[12] that a possible reason for academic authors' reluctance to publish in "electronic-only" serials is their concern for the long-term preservation of the academic record. Certainly, there is anecdotal evidence that such a concern stands behind some academic librarians' reluctance to move to an "electronic only" strategy for their serials collection. This has pushed some publishers of journals publicly to acknowledge that the development of an active policy relating to the deposit of their journals with national libraries is becoming an essential element of their long-term commercial strategy.

In this context, it is notable that major journal publishers in the Netherlands and Germany are already depositing electronic versions of their journals on a voluntary basis with their national libraries.

At this stage, it seems to us to be unlikely that in any other sector of the publishing market either authors or customers will be found to be so concerned about issues relating to long-term preservation; this is likely to be reflected in publishers' attitudes towards the deposit of many of their digital publications. While they may be entirely happy to deposit their publications (so long as their commercial interests are not damaged) they are unlikely willingly to take on substantial additional cost or effort in order to facilitate long-term preservation.

The adoption of standards for electronic publishing will be driven by other (market-related) factors. Any benefit that this may bring to the business of long-term preservation will be purely serendipitous.

Some specific examples may deserve exploration here.

A simple one relates to the adoption of standard identifiers (a subject which we believe will be of great significance to the future management of deposit).[13] The near universal adoption of the ISBN was driven by the commercial demands of the channel of distribution. A significant subset of the STM journal publishing community has recently found a similar commercial driver for the widespread adoption of the Digital Object Identifier (DOI) at the individual article level, in reference linking.[14] However, a recent conversation with an electronic publisher in another sector suggests to us that some publishers see common identification standards as being (potentially at least) at odds with their commercial interests. It will be a long time before that sector adopts a similar solution.

A more significant example may be the adoption of PDF. Despite fierce competition between proprietary page description languages during the 1980s, Adobe's success in establishing PostScript as the *de facto* standard in the graphics arts industries has been complete. The close relationship between PDF and PostScript allowed a few publishers quickly to adopt PDF as an electronic publishing platform very simply and at comparatively low cost. The "network effect"[15] rapidly created a community of producers and consumers that has increasingly expected PDF to be used for the dissemination of page-based electronic publications. Despite the claimed superiority of other formats, it is very notable that all but one of our respondents who are delivering page-based images are using PDF to do so (see Section 5.4)

Where information from multiple sources has to be combined into a single resource, the drive towards standardisation is strongest. The recent announcement of NewsML, an XML-based standard for managing and interchanging multimedia news, is an obvious case in point.[16] NewsML will allow the easy distribution of syndicated news in conformance to a standard XML DTD. For the time being, we can foresee no similar developments in other areas of

[12] We have been unable to find any published research that directly supports this.

[13] See Bide M, Potter E J and Watkinson A (September 1999) *Digital Preservation: an introduction to the standards issues surrounding the archiving of non-print material* Book Industry Communication for the BNBRF (freely available for download from www.bic.org.uk/digpres.doc)

[14] The CrossRef initiative: www.crossref.org

[15] See Shapiro, & Varian, H R (1999) *Information Rules* Harvard Business School Press for an extended discussion of the adoption of standards in the network environment; particularly relevant is the question of demand-side economies of scale (see pp 179 – 182).

[16] See www.iptc.org/NewsML.

content, since the need to interchange content has not proved strong enough to generate real enthusiasm for the adoption of standards.[17]

3.4 Publishers and metadata

The generation of bibliographic data by publishers has been notorious for its inaccuracy and inconsistency. Publishers have generally depended on third-party bibliographic agencies for the maintenance and dissemination of information about their products – and indeed have often re-purchased information about their own products.

However, recent developments are making it increasingly clear – to some publishers at least – that accurate and consistent metadata is an essential aspect of their business in the network environment.

One element of BIBLINK,[18] a European project that involved several deposit libraries as partners, tested the principles of the interchange and "improvement" of simple metadata records (based on Dublin Core) between publishers and libraries. Dialogue between libraries and publishers during the course of that project revealed that many publishers – particularly the smaller and "newer" ones – are becoming increasingly aware of the fact that their metadata is inadequate (or completely absent). They are actively seeking appropriate standards for the management and interchange of metadata.[19] Libraries involved in the project, aware of the proven difficulties involved in cataloguing electronic publications, made considerable efforts to raise the awareness among publishers of the need for "electronic title pages" to be included in their publications (carrying basic cataloguing metadata).

In the electronic publication of academic journals, publishers realised from the outset that article level metadata would be an essential element of any online delivery service; as a result, even where the primary delivery format has been PDF (which has been the case for the majority of ejournal publishers), article "headers"[20] have consistently been provided in SGML.[21] These now have the benefit of being produced as part of the production workflow and should therefore be consistent with the published document.

The most significant current driver of metadata developments in the book publishing community is the pressure from online booksellers for more extensive descriptive information. The development of the EPICS data dictionary[22] and even more so its subset, ONIX International,[23] have been strongly motivated by the need to provide information in a consistent form to the online bookselling community (a channel that is now of real significance to publishers).

[17] We were involved several years ago in an attempt to drive a standard SGML DTD for the interchange of journal "headers". (Bide M and Moore K (1996) *Electronic Tables of Contents (EToCs) for Books and Serials: standards for structure and transmission* Book Industry Communication.) The resulting SSSH DTD (see www.bic.org) built on the earlier work of the "Majour" initiative The standard has not been widely adopted, although one of the publishers to whom we spoke still uses the Majour DTD for their headers and we believe that many – perhaps most – journal publishers have based their own "header" DTDs on Majour or SSSH. We have not heard and do not believe that the SSSH is any way deficient. There is simply inadequate commercial demand for strict adherence to an externally generated standard to overcome the advantage for each individual publisher in developing its own local variants.

[18] http://hosted.ukoln.ac.uk/biblink/

[19] The difficulties of using Dublin Core as the basis for an extensible metadata set is well expounded in Lagoze C (June 2000) *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience* Presented at Seminar on Metadata organized by Archiefschool, Netherlands Institute for Archival Education and Research available at http://www.ncstrl.org/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1801

[20] The article header includes the abstract.

[21] Although, as we have already discussed (see footnote 17), publishers tend to use proprietary DTDs for their headers.

[22] The EDItEUR Product Information Communication Standards (see http://www.editeur.org/epics.html)

[23] See http://www.editeur.org/onix.html for more details

ONIX International provides a data dictionary and a standard XML DTD as a mechanism for the electronic interchange of a limited (but nevertheless rich) bibliographic metadata record. Although the initial release is focused on conventional print on paper books, future releases are expected to cover not only ebooks but also other media.

The extent to which all publishers will prove able to manage their metadata to a sufficiently high standard to meet market requirements remains to be seen. However, it is notable that a growing number of publishers, certainly in the US and the UK, are making significant systems investment in this area. The adoption of the ONIX International standard appears to be moving ahead with surprising speed in the US and the UK; and there is a real possibility that it could move towards rapid global acceptance in the publishing industry (although we recognise that the maintenance of high quality metadata demands much more than the publication of technical standards).

We believe that publisher-generated metadata will be particularly significant in the whole area of deposit of electronic publications; otherwise the conceptual and technical difficulties of cataloguing will be extremely difficult to overcome.[24] To what extent does ONIX International meet libraries' cataloguing needs?

A forthcoming conference presentation by Priscilla Caplan[25] suggests that libraries should

> *"…begin thinking about basic bibliographic metadata as a commodity, produced and exchanged by a number of communities in order to serve a number of purposes.... We will soon be in an environment where most metadata is exchanged in XML: the publishers have already adopted it, and library systems are moving in that direction. In this context it makes very little sense to think that libraries, publishers, booksellers, distributors and vendors will all be creating incompatible, non-reusable bibliographic metadata…I do urge librarians to take a serious and objective look at the metadata schemas emerging in the publishing community with the long-term goal of maximizing the interchangeability of data.".*

Caplan goes on to suggest that libraries should "work proactively with publishers to establish enough commonality between respective rule sets to allow meaningful exchange and reuse of metadata."

This co-operative approach was also endorsed during a workshop on digital preservation and deposit held in London in July 2000 (involving publishers and the UK copyright libraries).[26] For example, it became apparent during the conference that the technical data that (at least some) publishers are gathering for the management of their own internal content repositories is essentially identical to that required by libraries for the long-term preservation of the same resources. As far as we can see, there is nothing to be gained from recreating essentially identical data.

The EPICS data dictionary is designed to be extensible in ways that will encompass the necessary technical description of resources and it is anticipated that co-operative work between publishers and librarians (under the auspices of EDItEUR) will enable the development of such an extension.

However, it is important to remember that only a small minority even of large publishers currently have sophisticated digital content repositories; most are continuing to outsource their content management requirements (and in the opinion of many of those to whom we have spoken will continue to do so for the foreseeable future). This can only make the protocols for metadata exchange more complex to manage.

[24] See Bide M, Potter E J and Watkinson A (September 1999) Op Cit.

[25] At the Library of Congress *Conference on Bibliographic Control in the New Millennium* to be held in November 2000. Presentation available at http://lcweb.loc.gov/catdir/bibcontrol/caplan.html

[26] See www.bic.org/digpres for copies of presentation and the workshop report.

4 Publication media

4.1 Offline publications

In the offline category, we will consider only CDROM publications: only one of our respondents mentioned any other offline media in their publishing.[27] Seven (50%) of our respondents are publishing "stand alone" CDROMs.

For us, perhaps the most striking response to our questionnaire is the very rapid decline in interest in CDROM publishing among most of the publishers in our sample. Several of those we spoke to (particularly but not exclusively professional publishers) have established very satisfactory CDROM-based publishing businesses (which they expect to maintain for the foreseeable future) and some publishers use CDROM as an adjunct to online publication (either in hybrid publications or for the retrospective "archival" publication of offline journals).

Most did not expect to be publishing many if any *new* products (as opposed to continuing publication of existing products) on CDROM in future.[28] Increases in the available bandwidth[29] are simply rendering this style of publication, dependent on a physical supply chain, unnecessary and rather clumsy. Most publications previously published on CDROM seem likely to migrate primarily to online publication, although some migration to ebook publication must be a possibility.

This decline in offline publication[30] seems to us in general terms to represent good news from the point of view of libraries managing digital deposit. Although protocols for the physical deposit of offline publications may be a great deal easier to manage, the process of managing the long-term preservation of their content seems to us to be likely to be rather more complex.

This is partly because of the lower degree of standardisation of their content, when compared with web-based products.[31] CDROM products (unless they are web-browser based) deliver their content through proprietary user interfaces. Our respondents named a number of such interfaces; of these, only FolioViews and DynaText received more than one mention. Many, indeed most, CDROM user interfaces are specific to a particular product.

All of these CDROMs are typically Windows products; only four of our respondents produce products for Macintosh (and none of these exclusively for the Mac); two specifically mentioned that some of their products also run under UNIX.

The other complication with the preservation of offline publications is that, frequently, the content is encrypted. Publishers themselves, because they do not manage the production of their own CDROM publications, often do not have easy access to non-encrypted versions of the content.

4.2 Online and hybrid publications

All fourteen of our respondents publish either online or hybrid publications. All twelve of the larger publishers publish products that are entirely online. Seven of the fourteen publishers (including both smaller publishers) publish hybrid products. Among almost all the publishers we spoke to, the trend is again away from hybrid products towards pure online delivery. As

[27] One professional publisher still has a best selling subscription product that is distributed on diskette.

[28] This agrees with what we have found elsewhere. See Bide M, Kahn D, Max-Lino R and Potter E J (February 2000) *The Scale of Future Publishing in Digital and Conventional Formats. A Report to the British Library Policy Unit* available from http://www.bl.uk/proj/concord/otherpubblpu.html

[29] The UK's largest domestic ISP, Freeserve, is reported as forecasting that **all** of its 2 million domestic subscribers will have migrated to ADSL within four years (reported in *The Independent* 29 August 2000), as the cost falls precipitously.

[30] One publisher told us that some library customers have online content delivered to them on CDROM for local mounting of the data (and that they would "strongly prefer" not to have to do this). However, this is not "offline publishing" but simply a matter of the convenient management of the distribution of online publications.

[31] In the absence of strong network effects.

we mentioned in respect of the offline publications, the requirement to deliver “high band width” content offline is seen to be of decreasing importance as available bandwidth increases.

All the online publications in our sample are delivered using standard Web browsers for their user interface; many, though, require users to have additional software for access to some content. Commonly, this is a requirement for the PDF Reader plug-in, but the lack of standardisation of formats for sound and audiovisual means that a variety of interfaces are needed for these elements (see Section 5.6). One publisher in our sample is also, for example, using a specific ASCII-based format for rendering chemical structures.

Publishers often provide supplementary material in their online content; it appears that the commonest file formats for such material are mainstream Microsoft applications (Access, Excel and PowerPoint were all mentioned).

Among academic publishers, there is very rapid growth in the publication of author-supplied supplementary material, alongside journal articles; this is seen as a significant value added of electronic publication and is typically published in whatever file format the author may provide. Access to this supplementary material therefore requires the user to have a copy of the specific application. While these may most frequently be in the form of research data, published as mainstream Microsoft application files (Excel, Access), this is not controlled. For example, two publishers mentioned that raw PostScript was used for some supplementary material. Two publishers mentioned TeX (or LaTeX), and we are aware of audiovisual materials also being submitted for publication in this way.

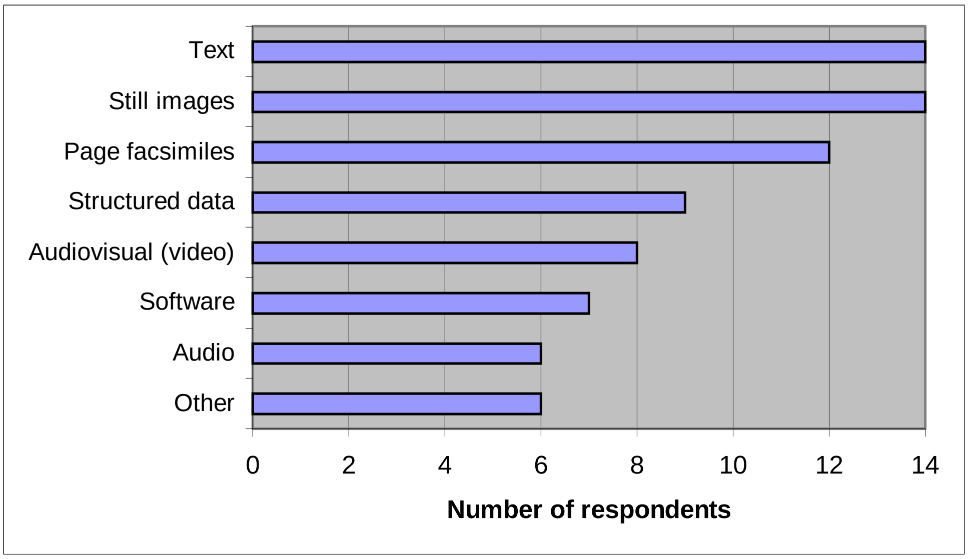
Standardisation of this supplementary material is therefore only achieved only to the extent that standard applications are adopted within the author community at large.

Security of content in online applications is usually achieved through access control, rather than encryption. Although the content may be inaccessible to Web harvesting techniques (see Chapter 7), it is very much easier for publishers and libraries to manage deposit processes with essentially unprotected content.

5 Content and content formats

5.1 Content types

The following diagram indicates the content types our respondents include in their products:



Unsurprisingly, all fourteen include both text and still images. Perhaps more surprisingly, twelve of the fourteen include page image facsimiles.

Structured data is the next most common content type. Video is included in products by more publishers than audio (only one publisher includes audio but not video). None of the publishers in our sample is making extensive use of either audio or video.

5.2 Content formats – text

Again, unsurprisingly, HTML is the common factor – all of our respondents deliver text in HTML to end-users. However, there is a distinct difference in approach between different publishers, in the use of “hard coded” HTML. About two-thirds of the publishers in our sample are generating HTML “on the fly” from SGML or XML coded text.

This brings us to what was perhaps the next most unexpected result in our survey – ten out of the fourteen publishers have repositories of SGML-encoded text, a much higher proportion than we might have expected.[32] Online publishing has clearly pushed the adoption of SGML-based workflows in the publishing industry.

One publisher is also making extensive use of RTF (but no other publisher mentioned this format).

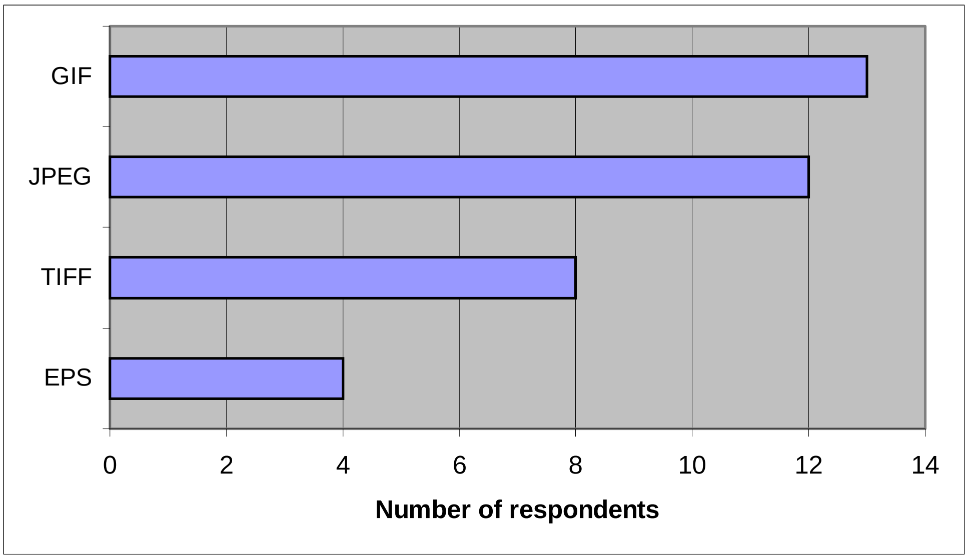
5.3 Content formats – graphics

The following diagram shows the use of different graphics formats by our respondents. It should be noted that most respondents are using more than one format.

EPS is used as an intermediary format, not for final delivery to end-users.

One respondent mentioned to us the possibility of using SVG (Scalable Vector Graphics format) for future online products.

[32] We also suspect that this is a much higher proportion than would have been found among a randomly selected sample of publishers.



5.4 Content formats – page facsimiles

All except one of our respondents who are delivering page facsimiles to end-users are using PDF to do so. The odd one out uses EPS.

One publisher is also delivering PostScript pages to a specific scientific community that has expressed a preference for this format (although it seems likely to us that this will prove only a temporary anomaly). Two publishers told us that their supplementary material also contains raw PostScript.

5.5 Content formats – structured data

As we have already discussed (see Section 4.2), structured data is usually provided in either spreadsheet or database formats (typically Excel and Access, although one publisher also mentioned Lotus 123).

Some online structured products are being held in proprietary database formats, with results of queries converted on the fly to HTML. This sort of product is typically dynamic (see Section 2.1) and therefore creates rather different challenges relating to deposit and preservation.

5.6 Content formats – audio, video and multimedia

There was very little consistency from our respondents in terms of their approach to formatting multimedia content. One specifically mentioned attempting to avoid proprietary multimedia solutions (like Shockwave) through the use of Dynamic HTML and JavaScript but this does not seem to be a uniform approach.

For audio, the formats that were mentioned by respondents were MP3, WAV, RealAudio. WAV was most frequently mentioned.

For video, respondents are using QuickTime, RealVideo and Lotus QuickCam – QuickTime was most frequently mentioned.

One respondent mentioned that they are using VRML.

5.7 Software

Software is most frequently included in computer science products. These are often (but not invariably) supplied on CDROM. The software may be provided either as an executable or as source code (sometimes both).

5.8 Other

Respondents mentioned various other types of content. These include: Microsoft Word templates; PowerPoint presentations; TeX for mathematics.

A number of publishers control the presentation of their content through proprietary stylesheets. A small number of publishers include fonts with their content, but most appear content to rely on default fonts available on end-user machines.

5.9 Dependence on middleware

A large number of products as delivered are dependent on middleware of one sort or another. Most commonly, this provides SGML/XML to HTML conversion. Search engines are also essential components of products, and these are often provided by a hosting intermediary rather than by the publisher.

One publisher of educational content also mentioned the significance of third party educational middleware. Their content is developed as courseware for integrated learning systems (WebCT and Blackboard were both mentioned). This content would be largely meaningless without preservation of the appropriate middleware.

5.10 Other formats

A small number of publishers are managing their published content in specific application files (like Quark Xpress or 3B2). However, there is general agreement that these are inappropriate for long-term archiving even by the publishers (let alone by third parties like libraries).

6 The ebook revolution

We have dealt with ebooks in a separate chapter of our report; the issues raised by ebooks are rather different from those raised by online resources (and current developments are even more immature and uncertain).

6.1 Should deposit libraries be interested in ebook technology?

Essentially, this question can only be answered by another question. Will ebooks have a major impact on publishing?

Although the timing of consumer acceptance of the reading of books on portable electronic devices may remain uncertain, there appears very little doubt (within the publishing community at least) that the ebook is set to have a fundamental impact on the dissemination of what we might call "book-like content" (see Section 2.1 for our working definition of an ebook).[33]

Some publishers have already made major commitments to the conversion of part or all of their backlist to ebooks; many of these (in the US in particular) are trade publishers, but the recent announcement from Taylor & Francis in the UK that they will digitise the whole of their backlist of 17,000 academic books cannot be ignored.[34]

However, despite much publicised initiatives from Stephen King and a few others, today the greater part of ebook activity is focussed on backlist titles (and therefore of little immediate interest from the point of view of deposit librarians – the titles being converted for ebook publication are already included in library collections in print-on-paper form).

Several publishers have told us that this immediate emphasis on the conversion of backlist is not accidental. There is so much uncertainty about the impact that ebooks may have on frontlist sales that publishers are unwilling to add to the considerable uncertainty that already exists in their sales projections (with the possible impact that this could have on increasing their inventory of unsold printed books). Conversion of backlist provides the potential of revenue from an otherwise dormant asset, without creating significant risk (not least because those who wish to sell the technology are for the time being often heavily subsidising the process of digitisation).

However, if ebooks become a significant medium for the consumption of backlist titles, it can only be a matter of time before market demand drives publishers in the direction of publishing their frontlist in ebook formats. "ebook-only" publishing may currently be largely confined to what many dismiss as an enhanced form of vanity publishing in the US;[35] this is unlikely to remain the case for long.

For this reason we believe that libraries cannot afford to ignore developments in ebook publishing, but should begin to think about the implications for deposit. While it may be several years before their impact becomes very significant, we believe that it is unlikely that ebooks will not be having a major impact on publishing by the end of this decade.

6.2 Are our respondents interested in ebooks?

Half of our twelve larger publisher respondents (representing a wide cross section of the different publishing sectors, with the exception of professional publishing) are either already

[33] A new study by Seybold Research shows that a high proportion of people are willing or eager to read e-books (66%), but they want them to be free. http://www.thestandard.net/article/display/0,1151,18062,00.html. Others suggest that the technology is two years from being ready for mainstream adoption – a view with which we are inclined to agree (see http://www.ohio.com/tech/news/docs/007948.htm). However, two years seems to us to be a very short time in the general scheme of things.

[34]See http://www.tandf.co.uk/statement.html. One of our respondents told us that a similar announcement could be expected shortly from their company, involving another technology partner.

[35] But see Max D. T. "No More Rejections" *The New York Times* on the Web July 16, 2000. Not all the books being self-published in this way are necessarily of poor quality – although this article focuses on books published by print-on-demand, the issues are the same.

publishing in ebook formats or expect to start doing so shortly. However, only one has already published anything *solely* in ebook format; and of the other five, three have no plans to publish only in ebook, and the other two say they regard it as premature to do so in 2001.

In terms of their commitment to ebook publishing, the five publishers already involved range from the very tentative and experimental to the fully committed – one of them is expecting to have their complete (US) trade frontlist in ebook (in parallel with print publication) in 2001.

The other publishers told us that they currently have no plans for publishing in ebook formats.

6.3 Ebook content delivery formats

The same "content and form" division that we discussed in our introduction (see Section 2.2) is also being played out in the ebook arena. ebook devices are divided between those which use content marked up for form (broadly speaking in XML, based on the Open ebook standard DTD – OEB)[36] and those which use page-based content (in PDF). The following table sets out the formats used by each device and the current distribution channel for these formats in the United States.[37]

Primary Format	Derivative Format	Current distribution channel (in the US)
OEB	Microsoft Reader	bn.com Lightning Source
	Rocket eBook (originally HTML)	bn.com Powell's, etc.
	SoftBook	SoftBook Press
	Peanut Press	Peanut Press
PDF	Acrobat Reader	Adobe PDF Merchant
	Glassbook Reader[38]	bn.com Glassbook, etc
	Print-on-Demand	bn.com Lightning Source Replica Books Sprout
	SoftLock Systems	SoftLock Systems

One of our respondents referred to the OEB standard as "little more than HTML expressed as XML". He went on to say that, as a result, the existing version of OEB, although theoretically extensible, is more suitable for novels than for more complex text. More complex ebooks can currently only be delivered using PDF-based formats.

The OEB standard is also undergoing revision, and another of our respondents expressed concern about version control (with the implication that there might rapidly be incompatibility between different versions).

[36] See http://www.openebook.org

[37] We are indebted to Ken Brooks of Barnes & Noble for providing us with this table.

[38] The acquisition of Glassbook by Adobe has just been announced. Adobe have also announced an ebook reader for the Palm OS, which apparently will "reflow" text for the small format screen from standard page size PDFs. This could prove to be an important step forward in the development of a more universal format for ebooks.

It is also important to recognise that each of the different derivative formats are different file formats, to a greater or lesser extent. Although we expect to see the different devices becoming more conformant with standards as time passes, there are significant divergences today. Publishers who wish to provide content to multiple devices need to maintain multiple electronic versions of the same title. This is even the case with PDF-based formats, where the text layout for one device may not be suitable for another (and therefore pages may have to be regenerated several times).

6.4 Producing ebooks

For the time being at least, ebook production has not been embedded into publishers' standard production workflows. ebooks today are typically being produced by scanning of printed products – we are told that this is often the case even for newly published titles (let alone for "deep" backlist).

Publishers' workflows for books in the print on paper environment are typically much less standardised than they are for serial publications. Except in the case of major structured reference works (for example encyclopaedias or dictionaries), neutral coding of text is rare or non-existent.

It is likely to take some time for ebook production to become a settled part of publisher's production workflows. Meanwhile, production workflows will remain somewhat ad hoc. It has also been suggested to us that vendors of tools for production of ebook formats may have a strong commercial interest in obstructing the adoption of stable, capable, neutral formats (which may further delay the process).

6.5 ebooks and other content types

For the time being, ebooks are generally restricted to fairly straightforward text and still images. We would regard it as unlikely that this will remain the case for very long. The OEB standard allows for arbitrary external technology to manage other content types (including complex text), and it may be some time before there is any general agreement on the optimum file formats. In the meantime, in the same way as we have seen in online resources, there will almost certainly be many different proprietary approaches.

6.6 ebooks and metadata

The OEB "package" incorporates a simple Dublin Core (DC) metadata record in Version 1. We understand that is has been recognised that the DC set is not adequate for the purpose, and a significant revision of the metadata structure is expected in future releases.

6.7 ebooks and encryption

An essential element in the marketing (to publishers) of all ebook devices is security, because of the "distributed" model of distribution (very different from the online model, where instances of the content are typically limited in number and controlled). The content on ebook devices is encrypted to prevent it being passed along to other users. There is an industry grouping, EBX,[39] which is defining a standard approach to rights management and encryption.[40]

The different approaches used for encryption need not detain us here. However, the fact that the content is encrypted (and may be tied to a specific hardware device, or to a specific user) does create a particular problem for libraries, not least in managing long-term preservation.

If ebooks are to be deposited with libraries, it will clearly be advantageous if these are non-encrypted versions. However, publishers will be very reluctant to allow libraries to provide any public access to unencrypted versions.

[39] http://www.ebxwg.org/

[40] The EBX working group mission statement begins: "The Electronic Book Exchange (EBX) Working Group develops open, freely available, and commercially viable standards for the secure transmission of electronic books (e-books) among rights holders, intermediaries, and users. EBX addresses such issues as the purchase, sale, lending, giving, printing, subscribing, and licensing of electronic books." It seems to us that the work of EBX probably has little relevance for the deposit of ebooks.

6.8 Managing ebook content

Only a small number of publishers are currently planning to manage their own ebook content in house. Most will be happy to see their ebook content managed by outside vendors (see Section 3.2).

Since this means that publishers will not be in direct control of their content, it is likely to complicate the process of deposit, particularly deposit of unencrypted versions of content.

7 Publishers and content management – an explanatory note

We have made a number of references in the text to publishers' attitudes to content management, and it seems appropriate to draw together here some explanation of the issues involved.

7.1 Historic attitudes

Traditionally, as we discussed in Section 3.2, the great majority of publishers have not been closely involved with the technical aspects of their production processes. Although most books and journals have been produced using digital technology for typesetting for over twenty years, the management of the digital files used for their production has not – until very recently at least – been a high priority for publishers: the cost of managing the content has not been matched by any measurable commercial gain from doing so.

To the extent that anyone managed the digital content, it was seen as a job for the outside suppliers – typesetters – on whom publishers depended (and, for the most part, *still* depend) for their text preparation.

The "DTP" revolution of the mid 80s did not, in most instances, lead to a major change in the approach of book and journal publishers. Some publishers – particularly those preparing highly integrated educational textbooks, or books with similar design requirements – successfully brought their text formatting in house.[41] But most (often following expensive but ultimately unsuccessful experimentation) continued to depend on external suppliers for the majority of their requirements. They simply found that managing the text preparation and formatting processes in house was too expensive and too difficult to manage.

The decision not to bring the work in house was supported by the collapsing price of buying-in services from outside vendors, whose adoption of much lower cost, PC-based systems (alongside much lower keyboarding costs, either through reuse of authors' disks or through "offshore" keyboarding in developing countries) has brought about extraordinary cost reductions in typesetting in real terms.[42]

Of course, there was some waste in this system. Often, when new editions had to be typeset, files for the old edition could not be found or were incompatible with a new generation of equipment. But the overall system itself was efficient – the difficulty of guessing which files you might need again would have made it essential to manage *all* files; and the cost of doing this was prohibitive.

7.2 The arrival of electronic publishing – the ejournal

Publishers therefore arrived at the dawn of the electronic publishing revolution with little or no experience of managing their own content.

There is a fund of stories about journal publishers who believed that they had access to their digital archives because their typesetters held years of typesetting tapes. Unfortunately, these tapes proved entirely inappropriate for conversion.

As a next step, PDF appeared to offer a simple solution to the problem – it could be produced easily and cheaply enough as a by-product of the current print production process (and brought with it the advantage of maintaining the print-on-paper product model as the primary business model).

There was a rapid recognition among journal publishers that they would have to produce SGML-encoded "headers" (see Section 3.3), but this was typically tacked onto the side of the existing production workflow (often extremely inelegantly!).

[41] As, of course, have virtually all magazine publishers.

[42] In the UK, typesetting (in cash cost per page) is no higher now than it was thirty years ago.

One or two journal publishers (including some of the largest) recognised very early that full text SGML would, in the end, prove to be essential to journal publishing and moved early to full SGML workflow for their journals. However, the difficulty of managing SGML workflows on this scale is tremendous, the costs involved high, and the quality assurance issues raised difficult to overcome (particularly where the process involves multiple third party suppliers, as it typically does).

There are many difficulties in managing content from multiple sources. These can be as simple as enforcing file-naming conventions and ensuring control of fonts and character sets. Most journal publishers set out on this process around a decade ago, and have learned many very difficult lessons along the way.[43]

Some publishers have worked their way fairly successfully through these issues – others have still barely started on the process of moving to "full text" SGML. It will be several years before *all* mainstream journals are available in a neutrally coded form, and the content sufficiently well managed for its straightforward transfer from publisher to deposit library.

Most of the major journal publishers are now managing their completed content in house and are also hosting their own content online. (Very few are managing a high proportion of their own text preparation in house). Many of the smaller journal publishers depend entirely on outsourcing for text preparation, content management and online hosting. This is unlikely to change, since few smaller publishers will be able to justify (or afford!) the systems investment required.

It is extremely hard to predict whether the trend in future among larger publishers will be a return to outsourcing or the continued development of ever-greater technical resources in house. Publishers have in the past tried to outsource as much of their operations as they can, and we believe that ultimately this may prove to be the overwhelming instinct.

7.3 The digital revolution and books

Journal publishers had a significant advantage over their book publishing colleagues – the relative uniformity of their content. Journal publishers have found very considerable problems in arriving at uniform (in house) DTDs to cover all of their journals, and have often found this to be unmanageable; a uniform DTD to manage all of a publisher's book output is unlikely to be attainable.

Publishers have not therefore looked to manage their books (for the most part) in SGML, but some of the very large publishers have recently started to look instead at ways of managing their content in the multiple formats that they already have – in the word processor format supplied by the author, in the native file format of the formatting software used by their typesetter (perhaps Quark Xpress, but perhaps something rather more obscure), in PostScript, in PDF. Available technology for managing digital content on this scale, with all the implicit version control for example, is expensive and mostly ill-suited to publishers' needs. The technology comes either from the print industry or from online content management – whereas publishers, for the foreseeable future, require a combination of the two which proves very difficult to achieve.

While a small number of large publishers have now taken the plunge into managing their own book content repositories; most have looked at the scale of the task and backed away from it.

Now comes the challenge of managing yet more multiple formats, for ebook publishing. As we have seen, while this may theoretically involve holding only a single master format and generating the content for different application formats as required "on the fly", it seems likely to be more complex than that, at least for the immediate future.[44]

[43] It should be noted that many professional publishers got to the business of neutrally-encoded content management rather earlier, presumably because of earlier market potential for electronic products.

[44] Although proprietary technology (based on proprietary file formats) from at least one vendor – Versaware – seems to promise exactly that.

In these circumstances, it is perhaps unsurprising that many book publishers appear to be making the decision to outsource their content management, at least for the time being. It may well take five years – as it did for the journal publishers – for patterns of ebook publishing to settle down sufficiently to allow many book publishers to make the major investments in content management that their journal publishing colleagues have now made. Even then, they may choose to outsource.

As we have earlier discussed (see footnote 10) new publishers, established to take advantage of electronic publishing from the outset (and with no print-based management legacy) may make the transition more easily. But the potential for a proliferation of new, small publishers will lead to its own problem from the viewpoint of deposit of electronic publication.

7.4 Implications for deposit libraries

There is a plethora of companies involved in the processing and storage of published content. These include (as examples only): typesetters; specialist online hosting services like CatchWord; content-management technology companies like Versaware; ebook and print-on-demand intermediaries like Lightning Source (and Barnes and Noble); and many more. The roles they play and the services they provide are different for different publishers. The processes and relationships are far from standardised at this point.

The services provided by these companies in electronic publishing may have a direct impact on the content that might be delivered to deposit libraries; on file formats and protocols, for example; and on metadata relating to the content. However these companies are typically operating simply as agents of the publishers and other rights owners (including, of course, a growing number of self-published authors). It is neither appropriate nor possible for these companies to develop direct relationships with deposit libraries (any more than it would be appropriate for publishers’ conventional production suppliers – typesetters and printers).

Deposit libraries will have no option but to deal with these suppliers through the publishers – any attempt to establish a direct relationship would be regarded as entirely unacceptable by publisher and supplier alike. However, it is vital to recognise that there may often be greater technical expertise the ends of the chain than in the middle, which may create some barrier to effective communication.

8 The public Web

8.1 Publishing and the World Wide Web

When we speak of "the public Web", we mean that part of the Web that is publicly and freely accessible to any user of the Internet. Despite extraordinary growth in many other published media during the 1990s, the World Wide Web represents the easily largest single explosion in the volume of publishing in the last decade, with estimates that the total number of pages on the "public web" worldwide has now passed one billion.

Most of those who post content to the Web probably do not regard themselves as "publishers" – although we would contend that they most certainly are, even if their published output occasionally reveals some lack of real publishing expertise! This explosion of publishing activity has been built on the rapid and universal adoption of a standard, and proves (if any proof were needed) the power of the network effect on the adoption of standards.

8.2 Content standards on the Web

Experiments in "Web harvesting" for archiving in both Scandinavia and the Netherlands have demonstrated that the overwhelming majority of files harvested have been in common formats. This would suggest that this is also true for the public Web as a whole.

Year of collection	**1998**	**2000**
Country of collection	**Sweden**[45]	**The Netherlands**[46]
Sample size (number of files harvested)	179,784	79,879
File type		
HTML tagged text file	56%	66%
GIF image file	20%	24%
JPEG image file	10%	6%
Plain text file	9%	2%
All other file types	5%	1%

Results broadly similar to those found in Sweden were also found in Finland in 1997.[47]

Without more data than is available to us, we cannot immediately suggest reasons for the marked variance in the data between 1998 in Sweden and 2000 in The Netherlands. It may suggest some difference in sampling methods (certainly The Netherlands sampling was not entirely random). It is also impossible to extract from this data the "average complexity" of a Web page – although clearly there must be a very high proportion of Web pages that are simply HTML tagged text.

It is unsurprising to us that the public web (which is primarily content which is not being published for direct commercial gain) should be relatively simply structured. For obvious reasons, harvesting operations of this type are unable to access content that is protected, or content that is dynamically served.

[45] http://kulturarw3.kb.se/html/state-98-01-26.html

[46] Lex Sijtsma, Koninklijke Bibliotheek (July 2000) Personal communication

[47] Juha Hakala, Helsinki University Library (July 2000) Personal communication

8.3 Digital preservation and the Web

The low cost of becoming a "publisher" on the Web has been the major facilitator of all this activity, and has significantly lowered the barriers to entry. This lowering of the significant investment barrier to becoming a publisher raises a number of new problems for all users of information – not least the deposit libraries. In the past deposit libraries were content to allow the commercial hurdle of publication to provide a primary filter in determining what should and what should not be deposited. With the Web, that primary filter is no longer in place.

Most of these "publishers" are not very prolific: in the Netherlands harvesting exercise to which we have already referred, over 80% of the domains harvested had fewer than 10 pages of content – and fewer than 1% had more than 500. Much of this "publishing" is at the level of personal home pages, which are almost certainly of only passing interest – the ephemera of the digital age.

However, some of what is published has lasting value – even though it may be distributed for nothing. Because of the negligible cost of distribution, the Web provides a mechanism for the publication of some content that (because of its specialist interest) would never have justified publication in print on paper form, including for example doctoral theses and a growing number of conference proceedings. At the same time, it allows Governments to provide free access to information that was previously difficult (or expensive) to find. It will not be at all easy for deposit libraries to separate that which has value from that which does not in the absence of the commercial "cues" which are available in the print world.

The primary issues concerning deposit and preservation of this content are not related to file formats and similar technical production issues; rather they relate to the amount of human intervention that is required in selection, identification and description of the content. Most of these publishers will not have heard of the term "metadata"; nor will they have given thought to issues of unique identification. There are other issues to be considered, like identifying the "boundaries" of a publication in a hyperlinked environment. Where does one document start and another begin? Can you only usefully preserve a document if you preserve all the documents to which it links?[48]

Although the purely technical issues relating to preservation of the Web may be difficult, particularly as the Web becomes more technically complex, it seems to us that finding answers to some of these other questions may ultimately prove to be more difficult.

[48] This problem applies equally to all online publications as hyperlinking between commercially published content becomes more commonplace, as exemplified by the CrossRef initiative (see footnote 14).

9 Conclusions

9.1 Reasons for optimism

In our research, we found a number of reasons why libraries with an interest in the long-term preservation of electronically published content might be optimistic about trends in publishers' production. These included:

1. The very rapid trend in the direction of full SGML/XML mark up of text, particularly among journal and professional publishers. It seems to us that SGML/XML mark up come closest to providing a format that is susceptible to long-term preservation of content.
2. The near universal acceptance of PDF for page facsimiles.
3. The small number of formats in use for still images.
4. The trend away from "offline" publication (which is very unstandardised) towards online and ebook publication.
5. The drive towards standardisation of ebook file formats.
6. The increased sophistication in some publishers' approach to content management, which means that their electronic content is being stored in a more consistent and coherent form than has been the case in the past.
7. The increased interest and investment by publishers in the generation of metadata, both bibliographic and technical.

9.2 Reasons for realism

This optimism needs to be tempered with realism.

1. SGML and XML, while apparently "neutral" in their application, are never entirely neutral. Publishers develop their DTDs to meet their own specific needs and workflows. Publishers using SGML amend their DTDs, if not frequently at least occasionally, to overcome deficiencies on one kind and another. This implies major tasks for libraries in managing SGML-encoded text, in terms of version control of text and the DTDs that accompany it.
2. It is highly unlikely that any time in the near future their will be a trend towards standardisation of DTDs even by publishers in the same sector, except where (as in the case of NewsML) there is a real market requirement to share syndicated content from a number of sources. We do not see this happening in the near future in any other sector.
3. At the moment, at least, the SGML and XML archives that are available do not constitute the "published product" – the product is dependent on middleware to deliver it to the end user. This middleware may or may not lie within the publisher's control. This situation may become a little easier as XML-capable browsers become standard, and publishers can use the full capabilities of XML.
4. An increase in the use of SGML or XML will prove to be a double-edged sword. Greater sophistication in a publisher's ability to "slice and dice" content is likely to lead to a proliferation of new units of delivery. Publishers are already talking about using their SGML/XML-based archives as a means of extracting more granular pieces of content for new kinds of commercial exploitation. Individual chapters might be sold, or even smaller chunks in combination with material from other sources. If this becomes a significant practice (as in certain markets seems likely), it will represent a new type of dynamic publishing model in which the content itself is static but the number of products increases by an order of magnitude.
5. Although a small number of formats may be used for text and still images, there is still a proliferation of formats used for other types of content. This is likely to remain

particularly significant in the case of author-supplied supplementary material to academic journals. The "standardisation" of the format of this content will depend on the standardisation of the applications used by the author community (or conceivably on the standardisation of file formats). Again, we do not anticipate that this will happen in the near future (although network effects suggest that it will happen in due course – the pre-eminence of Microsoft applications being a case in point).

6. We have long been promised the coming together of technology for managing "content and presentation" (for example, "structured PDF" has been talked about for a number of years), it seems unlikely to us that any significant breakthrough either in more effective "on the fly" formatting of structurally-marked-up content or more flexible management of page-formatted text is likely in the very near future. Publishers are likely to need to continue to publish in parallel forms for a considerable time to come. PDF will not disappear as a format for some time to come, and may indeed remain the format of choice for some types of publication.

7. Security mechanisms, particularly encryption, are likely to continue to create a barrier to straightforward long-term preservation, at least in ebooks.

8. We have been speaking to a group of technically sophisticated publishers (and to the most technically sophisticated members of their staff). Many publishers (even some larger ones) still do not employ personnel who would understand the issues raised in this report. This is unlikely to change; technical issues will continue to be outsourced. This will make communication between publisher and deposit library more difficult.

9. The problems relating to understanding of content issues relate equally to understanding of metadata issues. One publisher we spoke to recently told us that he was appalled by the continuing lack of understanding he had found (in both the US and the UK) of the significance of metadata to the future of publishing. Without better publisher-generated metadata, the management of deposit of electronic publications will be extremely difficult if not impossible

10. The new economics of publishing seem likely to lead to a considerable increase in the volume of publishing and the number of publishers. Many of those publishing will not recognise themselves as publishers. We believe that the significance of this for the deposit of electronic publications cannot be overstated.

11. Even in a straightforward technical survey of this kind, publishers continue to express some reservations about the deposit of electronic publications with libraries, particularly with respect to the provision of access. It was suggested to us more than once that deposit should wait until the commercial life of the content has been exhausted. Very close co-operation will be required between librarians and publishers to allow digital deposit to work satisfactorily – and it may prove difficult to motivate publishers to become involved in such co-operation. The cost of compliance with library needs will have to be low to publishers – and the difficulty of attracting the attention of publisher's hard pushed technical staff to meeting library requirements should not be under-estimated. In general, there is good will but resources are very stretched.

9.3 Some final thoughts

In general terms, publishers are driven by the market place. With few exceptions (like metadata), publishers will not develop standards for themselves but will adopt the standards that are widely adopted within their community of customers. This is clearly observable among our respondents (in the use of TeX, for example, or the retention of PostScript for a particular community of users).

Both online and ebook publishing will see the "network effect" take greater and greater hold on standards. The market will enforce standardisation wherever lack of standardisation creates inefficiency for the end user. The Web provides a very good example of the extent to which *de facto* standards (like HTML) can come rapidly to dominate an electronic distribution medium. Despite the limitations of HTML, publishers of all types of content have found "work rounds” to enable them to deliver content in this format, because of its widespread market acceptance.

Inefficiency for the end user and inefficiency for the deposit library are not necessarily entirely congruent. However, the trends are encouraging.

10 Acknowledgements

I am extremely grateful to the following people who responded to our email questionnaires, spared the time to speak to us on the telephone or helped us in some other way with the development of this report:

Tony Hammond, Emma Parkinson	Academic Press
Ken Brooks	Barnes and Noble
Peter Hearn	BMA Publications
Celia Mindelsohn	Croner-CCH
Rufus Neal, Kevin Taylor	CUP (UK)
Laura Dorfman, Gita Kapila	CUP (US)
Juha Hakala	Helsinki University Library
Eamonn Neylon	Informata
Cliff Morgan	John Wiley & Sons Ltd
Lex Sijtsma	Koninklijke Bibliotheek
John Peacock	Macmillan
Julian Clayton	Maris Multimedia
Jonathan Ward	OUP
Claire Cameron	Pearson Education
Nick Perry	Pendragon Professional Information
Dr Thomas Rakow	Springer Verlag
Paul Riddle	Sweet & Maxwell
Martin van der Linden	Wolters Kluwer Academic Publishers

I gratefully acknowledge the assistance of Francis Cave (Francis Cave Electronic Publishing) for carrying out the substantive research for this report.

Without the input of all of these people, this report could not have been written. All errors of interpretation of what they have told me are, of course, mine.

Mark Bide
Frimley Green: August 2000