VIRTUAL STACKS: STORING AND USING ELECTRONIC JOURNALS

Michael Alexander

Paper presented at the UKSG 20th Annual Conference, Edinburgh, April 1997

Although some interim de facto standards, such as the Internet interface and Portable Document Format(PDF), are emerging amongst the ever changing systems for the storage and dissemination of electronic information, many issues, including interoperability, metadata description and persistence in document identifiers, remain to be resolved. However, an information supplier like the British Library cannot wait for total system stability, and it has set up a trial document store, using both a local datastore and working with commercial suppliers on access to external data sources. Such co-operation may produce the practical solutions which are viable in the longer term.

Michael Alexander is Document & Image Processing Manager at The British Library, Information Systems, Boston Spa, Wetherby, West Yorkshire LS23 7BQ

I begin my paper by reviewing some of, what I consider to be, the more significant e-journal developments which have taken place recently, in several different domains. I then look at how the British Library is approaching the task of handling electronic journals; both as a remote document supplier and as a national repository of last resort. Finally, I return to some of the technical issues concerning storage in, and dissemination from, virtual stacks of electronic material which are raised by the current methods of creating and formatting e-journals.

I should like to make it clear at the outset that my approach to this topic is primarily from the direction of data processing, albeit in a library environment; in other words a context in which I see a number of technical challenges to be met and overcome rather than a concern with the far wider general issue of serials control in an electronic situation.

It is often all too easy to forget, particularly if one is quite deeply involved in the application of digital and networking technology to our working life, how very recent the penetration of this 'new wave' of computing technology actually is. I mention this at the outset as a reminder to us all to be aware that even now the methods of creation, format and storage which have been, and are being, used for electronic publications, are far from definitive and final. We are going to have to go a lot further before readers are generally as comfortable with electronic forms as they are with that particular version of the codex, the printed volume.

I see the current creation and use of e-journal literature taking place in three main areas. Firstly, there is the area of what I will call 'self-publishing', particularly where it takes advantage of the access and delivery mechanisms which the Internet has provided. Secondly, there are projects which are involved in the retrospective conversion to digital form of , often very old, serials. This is primarily for increased access, but also to some extent for preservation reasons and also in the hope, pious rather substantive at the moment, that material thus digitised might form the basis of a new revenue stream. Thirdly, there is the move by established journal publishers to avail themselves of the Internet delivery mechanisms and also to try and pre-empt any possible future decline in use of the printed volume.

It is scarcely surprising that major publishing houses, for whom the creation, preparation and printing of publications is now largely an electronic activity, should be seeking to gain some further advantage from their electronic data by making it available in digital as well as in print form. Nor is it surprising that, increasingly, the format adopted for this process should be Adobe Acrobat's Portable Document Format(PDF). Adobe's PostScript printer language is in effect the standard for transfer of data for electronic type setting and PDF is a spin-off or variant of that product and can be produced from postscript files with no additional human effort. The other favoured format for the digital publication of journal articles seems to be the TIFF image format. In fact some publishers who now use PDF first produced their serial titles digitally in TIFF form.

An interesting variation on this is to be seen in those titles supplied via the CatchWord system. The documents are held on the system server in a proprietary file format and converted to a TIFF variant 'on-the-fly' when delivered to the customer for local printing. The common factor for these formats is of course that they enable the visual form of the printed journal to be electronically retained, while reducing the possibilities of tampering with the content. Unsurprisingly, publishers who have spent much time and effort on developing house styles for their serials, which help their 'brand image' would not wish to see those simply overturned in the digital environment. Furthermore, we all know how uncomfortable reading large chunks of text on screen actually is. So while articles remain constructed of text and fixed images alone, the likelihood is that recipients of the digital copy will print it off for their convenience at the earliest opportunity. I will return to the implications of this and the pros and cons of such formats later on.

For large research libraries back runs of serials can be both a blessing and a stumbling block. The existence of such holdings is what contributes to their status but at the same time the unit use relative to the storage requirement is generally low, creating a large overhead. Older material becomes more difficult to conserve and handle and can often be made available only on microfilm. And as we all know reading large amounts of text on a microfilm viewer is even more objectionable than reading it on a VDU screen. In the UK the Universities' eLib programme is funding a project called the Internet Library of Early Journals[1]. A number of important 18th and 19th Century serial publications are being scanned in twenty-year runs. The images are converted to text using optical character recognition and indexed using Excalibur Technologies' Adaptive Pattern Recognition Technology engine. On retrieval the scanned page images are presented for reading via a web browser in UK universities. Using similar methodology but currently conceived as a commercial venture on larger scale is the JSTOR project[2] in the USA . Whole backruns of journals are being scanned as TIFF images, indexed and stored on a central server with the aim of offering university libraries and similar institutions access to this central store. The emphasis is very much on producing high quality printed output at the retrieval station for the personal use of the researcher.

The British Library has in a small way also contributed to this by scanning from microfilm runs of English local newspapers from the late Eighteenth century Together with a simple index these images will shortly be available at the Library's Newspaper reading room in London for readers to see whether they prefer this method of access to the microfilm copies which they are currently offered. As with the journals published in electronic and print form simultaneously these projects aim to reproduce as faithfully as possible a copy of the original printed page and are therefore less new publications than a new form of surrogate, provided to meet growing access or conservation demands on existing material.

The term self-publishing, when used in the print domain, often carries with it an aura of small independent presses, alternative, radical, at times anarchic, challenging the establishment. Those feelings certainly inform the attitudes of many sites on the World Wide Web and for most of us anarchic is the adjective which springs most readily to mind, as we try to find our way through the Web! Indeed, the very term may embrace everything from the revolutionary pamphlet, through the parish magazine to the occasional output by a small learned society. As in print world, so in the Web environment the content and form may be just as rich and diverse. One of the best known current examples of an establishment institution using the WWW as a medium for its own publications can be seen in the Science Preprints series, produced by the National Laboratory at Los Alamos in the USA[3]. Increasingly other such organizations are following the same route for dissemination, either to a restricted membership, by the use of password protection, or by making the documents freely accessible on a public web site.

More interestingly, perhaps, is the advantage now being take in many parts of the world of improving technology and bandwidth in order to move away from the traditional text and fixed image content of printed journals to a wider use of multimedia contents. This area is still very much in its infancy and trying to answer some basic questions such as: "What is a multimedia journal?; "Will authors wish to write for it?"; "Will authors know how to write for it?"

Again, eLib has been sponsoring some very interesting projects in this area[4] and commercial publishers too are exploring the potential. Unfortunately, would-be publishers of such new ventures will for some time to come have to face the problem of providing useful and useable contents upon a still developing and unstable technology. All in all, self-publishing on WWW is becoming a recursive activity, possibly even a tautology, because as the Web becomes the accepted interface for access to digital objects, those involved in the other two areas of e-journal publication, mentioned above, are drawn into Web publishing as the obvious method of disseminating their products. As a result, within the Web environment, for good or ill, the established criteria of large and small, rich and poor, quality and dross, begin to break down. It is all just Web publishing. In fact, given that a serial is a publication in successive instalments and thinking about the way Web sites constantly mutate within a single continuing title, one has perhaps to start considering Web pages themselves as a new form of serial publishing.

How does the British Library stand in this time of turbulent change? Well, we are fairly confident that the complete and ultimate solution to digital document control is as unlikely to emerge at this particular moment as treatise on bibliographic control published in the age of Caxton could be taken as definitive. However we have to respond positively to the way the world goes, and the world is going digital - and nowhere is that more clear than in the area of remote document supply.

The UK Serials Group will of course be well aware of the British Library's long involvement in document supply. From early days as the National Lending Library for Science and Technology through to the launch last year of the Library's new service entitled "Inside" our serials holdings have constituted perhaps the most important part of the document supply collection. "Inside", based on both our collections and on our successful Electronic Table of Contents product proposes to offer a more rapid delivery service, tailored closely to the requirements of both individual and institutional customers. Of course, a more rapid delivery service inevitably supposes a networked form of document delivery, whether it be by fax, e-mail or using an on-line client/server approach, of which the web browser is one of the more obvious client mechanisms. It also assumes that the document is available in a digital form and for that reason the Library has entered into licensing agreements with a number of serials publishers to make use of selected titles in a digital document delivery environment.

At the moment this process is supported by a trial document store, cunningly entitled TEDS, the Trial Electronic Document Store. It utilises a document management application called Image 6000 which had been acquired for an earlier project. Because the store operates in conjunction with our Automated Request Processing(ARP) system and requires no human interface it is in essence very simple. Each article is received from the publisher as a single PDF file together with a bibliographic description. The bibliographic description is parsed and a simple index to the article is constructed . The index is then passed back to ARP for identification of requested articles. When the ARP system identifies a requested document as being in TEDS it passes the request through to the store. The document is then retrieved. At this point I would like to be able to say that the document is then transmitted electronically. Unfortunately the software to handle that aspect of document delivery is still in development and so the journal article is printed out and dispatched in the normal way. Later in the year we plan to be able to pass the document to the delivery system, called SIDS, for onward transmission to the customer.

This is satisfactory enough for journal articles which are held on site, but the process for retrieving articles held in remote stores is rather different This part of the application is still in development and we are working with the company CatchWord Ltd. to make use of published articles held in their store. The first part of the process is similar. We would receive bibliographic information about articles in CatchWord's store to which we had access and create index data about it. In this case the pointer to the article would not be a file name in TEDS but an URL for the article's location in CatchWord. When ARP passes such a request to TEDS it will retrieve the article, using CatchWord's proprietary software RealPage. This will convert the article to TIFF pages 'on the fly' and these will be printed out or passed to SIDS (Scanning Integrated with Document Supply) in the normal way. Whatever the mechanism for retrieval, however, the assumption at this stage is that the end-user is receiving a printed copy of a text-based article.

Particularly for an institution which describes itself as 'One Library on Two Sites', we are keen to maximise use of this material by also making it available on-line in our reading rooms. That desire would become a requirement if some kind of deposit process were to be developed, utilising the British Library's status as a national repository of publications. Many of you will be aware that the Library has been active for some time in promoting the case for the extension of legal deposit to non-print materials and it now seems likely that some form of deposit arrangement, either through legislation or by the development of voluntary arrangements, will occur and we need to start preparing ourselves for this.

There have already been some interesting and successful developments in this area, Elsevier's TULIP project[5] in the USA and Tilburg University in The Netherlands[6], being two of the best known. We have no interest in re-inventing the wheel and intend to draw all we can on the previous experience of others. At this point it should be remembered that we are dealing with developments that by no means affect serials alone. Within the context of the digital library e-journal articles become one set of digital objects among many others to be sought , identified and retrieved by end-users. Clearly, the very simplistic model I describe in TEDS would fail to meet that level of sophisticated functionality. Furthermore, the scale on which a library of the size of the British Library needs to address these questions suggests a level of resource input which is certainly beyond our current capability.

It is for this reason that we have opted to pursue the route of private finance initiative (PFI), through which to develop a large-scale digital library application We have already begun to seek a partner or partners to join us in a venture which would include the digital development of both our remote document supply and patents services and the digital conversion and exploitation of selected areas of the historical collections. The PFI is expected to take some two years to reach fruition and we clearly cannot stand by and do nothing while waiting for this to happen. Consequently, this year we intend to begin planning the design and implementation of much broader-based and functionally-richer storage facilities than those available using TEDS. This will enable us, we hope, to manage more effectively the digital material that we are already receiving or about to receive, either by purchase, license or deposit, and to facilitate reading room access to this material in addition to remote document supply.

The successful accommodation of these functions is less within the storage mechanisms per se, although issues like server speeds will have a role, but more in the way in which this digital material can be defined and described sufficiently to allow end-user to navigate to it within the digital environment. This whole area of descriptive data has come to be known by the name of metadata; unfortunately in my view because I think its a rather sloppy use of the term but I fear we are stuck with it now. It actually encompasses a continuum of different types of descriptive data from data which is most useful to the human user, in the shape of catalogues, indexes, abstracts and other forms of finding aid, through the paraphernalia of the hypertext transfer protocol, to information about particular files, through name, format, creation date etc., which might be interpreted only by a computer. Definitions of this kind of data would be difficult enough in self-contained systems but the essence of the 'new' digital environment of the Internet is 'inter-operability'.

This is a rather unwieldy term for a very important concept. It implies that users will want and expect to retrieve digital objects from a variety of locations within the digital universe, and that a universal IT infrastructure to accommodate that desire will be in place. I know that this raises questions of rights and permissions which are by no means trivial but I believe that the opportunities presented by this technology are too great not to seek the satisfactory resolution of such problems. Nevertheless, I am grateful that I can leave that to others to carry out. If a system of "I'll show you my holdings, if you show me yours" is genuinely to develop then we had all better mean the same thing by that. Otherwise some people are going to be in for a bit of a shock when they get to where they think want to be. This brings me to some important technical issues which have to be dealt with before such a concept can have practical application in the long term.

Two central technical issues, if not the most important issues, for the long-term storage of and access to digital material are those of persistence, both persistent names and persistent locations, and of data integrity. I believe the two are directly inked to each other. We have all experienced the frustration, after having located some information on the Web and bookmarked it, of returning to it sometime later; only to receive the message that this item has been relocated to another site with a different URL, or even worse that bleak message, "File not Found. The requested URL was not found at this server", and being left stranded in hyperspace. I sometimes think that the term Uniform Resource Locator should be renamed Uniform Relative Locator. Moreover, even if one does retrieve the desired object what explicit guarantee is there, except perhaps a vague belief in the integrity of the particular site accessed that the object identified by the file name is exactly the same as the retrieved previously. We have already noticed that Web pages themselves mutate constantly - why not other digital objects?

Currently there are a number of apparently competing solutions to these problems but as I said at the start of my talk these are still 'early days' as far as ultimate solutions are concerned and I remain sanguine over the final outcome. The difficulty, for those of us involved in trying to build systems now, is what route we should chose and I am sorry to say I do not come here today with the complete solution. However, I am extremely interested in the work which Bill Arms and his colleagues at CNRI in the Sates are doing for the Library of Congress[7] and conveniently that brings me back to the particular question of identifying electronic serials material. As you are doubtless aware, serials publishers, in groups or individually, have for some time being grappling with the issue of producing a unique identifier for an article. The two codes which are prominent here are the SICI[8] code and PII[9]. Both are interesting attempts but possibly do not go far enough in facilitating their immediate use within the WWW environment.

The Association of American Publishers is addressing this very point by commissioning CNRI to develop what it calls a "digital object identifier (DOI)[10], which can incorporate codes produced either through SICI or PII and is intended to be a persistent identifier for that object within the digital environment throughout its life, regardless of ownership and other issues. Of course, it remains to be seen whether this proposal, and the rather elaborate 'registry' mechanism that it will require, will get acceptance from all the stakeholders involved in the electronic serials domain, but frankly I do not see substantial progress in the widespread use of networked serial data being possible without this, or some similar solution, being found.

Another important area of concern for long-term storage and use of digital objects is their format; that is , how the data content has been encoded for use in the digital environment and how that can be decoded for re-presentation to the end-user. Those of us who subscribe to electronic discussion lists know that this is a topic which can raise fierce passions, much of it I feel misguided and ultimately irrelevant. Open systems to support inter-operability is the preferred goal, but it has to be admitted that open systems, of the kind produced by international committees, do not have a good track record within the fast moving IT industry. Whilst it does try to avoid being locked into particular supplier solutions (although how many of us are not using Wintel PCs packed with Microsoft products?), it will usually go for the de facto over the de jure solution. This is not always bad; the Internet protocol itself is an example of the success of this approach. While I would welcome all digital documents produced entirely in strictly ISO-conformant SGML, I will not expect this to happen, any more than my colleagues who work in conventional library storage areas would expect all printed material to arrive in a single size, on acid-free paper and non-degradable bindings. We have to work with what we get.

I also think that it is reasonable to suppose that publishers will not adopt document formats which are unavailable to the bulk of their customers. In that sense, I do regard PDF as an open system in so far as it is based on PostScript which is a published standard, whose files could with some effort be reformatted over time, if this was required. Whether libraries would have the resource to carry that out, is another matter. As a systems manager rather than a miracle worker, I cannot guarantee to library colleagues that, for example, multimedia documents produced with proprietary authoring software to run in a particular computer environment will still be accessible twenty years hence. I know that library conservation staff in Europe and North America are looking at these kinds of issues but I really would urge library colleagues to take advantage of forums like the UKSG conference to engage in vigorous discussions of these issues with the publishing community.

Ultimately, both publishers and librarians face the challenge of inter-operability, which has commercial and resource implications for both groups. If the old adage "two heads are better than one" has any value, then perhaps together they can produce solutions which will be immediately cost-effective and practical to sustain in the longer term.

References

1. http://www.bodley.ox.ac.uk/ilej/
2. http://pjstor1.princeton.edu/
3. http://xxx.lanl.gov/
4. http://www.ukoln.ac.uk/elib/lists/ej.html
5. http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml
6. http://cwis.kub.nl/~dbi/cwis/col/bronnen/let.htm#Electronic journals
7. http://www.cnri.reston.va.us/
8. http://www.faxon.com/Standards/Z3956-SICI-Intro.html
9. http://www.aip.org/epub/piius.htm
10. http://www.doi.org/