DESIDOC Bulletin of Information Technology, Vol. 17, No. 6, November 1997, pp. 23-32

The Digital Library on Campus : A Perspective from a UK Academic Environment

David J Price*

Abstract

Whichever definition you wish to employ, the age of the Digital Library is upon us. There is much theoretical discussion about the nature of the Digital Library, but whatever form it comes to assume in the future, its foundations are surely being laid now. It is often remarked that there will be no place in it for librarians who will be displaced from the world of information to become merely curators of book museums. But this is not the experience of this librarian who is hearing more loudly now than ever before, demands from our users for information in digital form and an insistence that libraries should be exploiting IT on their behalf. To the academic community, we have been traditionally the gatekeepers to the world of information. They now see us as being effectively the purse-holders for information, and they are looking first to us to provide the information services they need for teaching and research.

The Follett Report [1], published in 1993, has had a tremendous influence in the UK in this respect. Its recommendations have led directly to the release of funds specifically for libraries which has provided a window of opportunity for institutions wishing to collaborate in digitization projects. The author discusses briefly some of those in which Oxford is closely involved.

Many forces have contributed to the new expectations of the library profession, but of perhaps most significance has been the accelerating growth of high bandwidth networks and Internet connectivity. At last it is becoming feasible for libraries to consider tackling the problems of the information explosion and shrinking budgets, by moving truly from 'holdings' to 'access' policies. There is also the demand that our information services should reach beyond library buildings to the users' desktops on campus and at home. As a practicing digital librarian who has had to grapple with the technicalities of providing integrated and ubiquitous information systems in the real virtual world, certain difficulties have become apparent which are common to most UK academic establishments and, I suspect, to those in Europe, the US and India too.

* *Head of Systems and Deputy Keeper of Scientific Books, Bodleian Library, Oxford, UK*

With respect to accessing quality CD-ROM and magnetic databases, mounted locally or remotely, the SilverPlatter ERL system has provided us with a powerful mechanism, especially with the development of WebSPIRS, which allows access to databases from standard WWW browsers, and provides support for Z39.50, which will allow access from Z39.50 clients.

The WWW has achieved a high integration of Internet sources and we are using it for many applications. But it has inherent limitations, in particular, the 'stateless' nature of the http protocol and the semantic poverty of HTML. We shall examine these issues in the light of new standards and technologies that are emerging, in particular, the IETF's work on Uniform Resource Identifiers, the Text Encoding Initiative and Sun's JAVA programming language.

1. INTRODUCTION

In my abstract of this paper I stated that the age of the Digital Library is upon us, but in a sense, it has been with us a long time. In the 1970s libraries began connecting to remote hosts to conduct online searches on digital—and usually commercial bibliographic databases. In the 1980s many libraries automated their catalogues and we saw wide-spread adoption of standalone CD-ROMs. Campus Wide Information Services appeared (usually with much content provided by libraries) and in the academic world it was generally possible to access remote CWISs and library OPACs. In the 1990s, we have seen CD-ROM networking, connection to suppliers and agents systems, document delivery, online electronic journals, the use for academic pursuit of email, email discussion lists and news groups, and much more.

Electronic sources bring with them special problems of management, many of them technological, which, in the world of books, librarians have not had to confront before. They range from acquisition problems to access restrictions, authentication, copyright, preservation, software and the user-interface. It is beyond the scope of this paper to cover these comprehensively, but I should like to bring to your attention what we consider to be some of the key issues as we strive towards developing the digital library, illustrating them with examples from our experience at Oxford. Firstly, I shall discuss about the digitization of existing library material. Next, the question of connectivity, the reality of the modern academic and the importance and problems of client/server computing. Then I will consider the integration of network resources and, in this respect, the importance of the World Wide Web. Lastly, cataloguing and indexing Internet objects, and the standards that are being developed to make this dream possible.

There is much theoretical discussion about the nature of the Digital Library but whatever form it assumes in the future, its foundations are surely being laid now. It is of vital importance that decisions we make concerning the standards we adopt, the systems we develop and the format in which we store our data, are well informed. It is often remarked that there will be no place in the Digital Libraries of the future for librarians who will be displaced from the world of information to become merely curators of book museums. But this is not the experience of this librarian who is hearing more loudly now than ever before demands from our users for information in digital form and an insistence that libraries should be exploiting IT on their behalf. To the academic community, we have been traditionally the 'gatekeepers' to the world of information. They now see us as being effectively the 'purse-holders' for information, and they are looking first to us to provide the information services they need for teaching and research.

2. DIGITIZATION

I shall make brief mention of several of the major digitization projects taking place at Oxford which exhibit the use of different techniques and different materials. Digitization does not happen cheaply and it is noteworthy

that all have been made possible only through the availability of external funds. Since 1989 the Bodleian Library has run a fund-raising campaign and some of this money has been allocated to developing IT. We have also been the beneficiaries of funds provided to joint European partners by the European Union's DG XIII Telecommunications and Libraries Programmes. But of greatest consequence in the UK has been the influence of the Follett Report whose recommendations led to the creation of FIGIT, the Follett Implementation Group for IT. The main thrust has been on collaboration between libraries to provide network access to their resources, enabling them to cope with the publication explosion and shrinking budgets. It is debatable the extent to which this will be achieved, nevertheless, digitization brings other benefits for classical library problems such as conservation, preservation, space and the ability of scholars to bring to the desktop images of manuscripts and documents from around the world for close inspection and comparison.

Under the auspices of FIGIT, thirty five Electronic Library (eLib) projects have been started in the seven programme areas of Document Delivery, Access to Networked Resources, Digitization, Electronic Journals, On-Demand Publishing, Training and Awareness and Supporting Studies. Later, calls have been made for proposals for projects in the areas of Hybrid Libraries (i.e. integration of digital and non-digital library information), Large Scale Resource Discovery (cross-domain searching) and Preservation of Digital Material, and these are in the process of being assessed. Another result of the Follett Report, was the allocation supported by HEFCE (Higher Education Funding Council) of more than #40m of non-formula funding to some 35 institutions over five years for the preservation, cataloguing and support of special collections in the humanities. It is intended to co-ordinate the results by the creation of a national OPAC of research collections and a national archive location service. As Derek Law has pointed out, we in the UK should be aware that these resources have a pump-priming nature and are only likely to be available for 3-5 years. Further, the Government has not provided new money, rather it has been top-sliced from the Higher Education budget [2]. In so doing, it reflects the importance that our leaders and policy-makers give to the development of digital libraries and their willingness to entrust this development to the libraries.

2.1 Celtic and Medieval Manuscripts

This is a one year pilot project with a four year continuation supported by HEFCE. forty four rare and fragile Celtic manuscripts (c. 16,000 folios) from the Bodleian, Jesus College and Corpus Christi College are being digitized using a Kontron ProgRes 3012 digital camera using 300-600 dpi and 24-bit colour. JPEG compressed files of about 800Kb are being made available via the WWW with thumbnail images and descriptive cataloguing derived from existing printed sources. The continuation project comprises Medieval manuscripts from the Bodleian Digby Collection and 70 manuscripts each from six Oxford colleges. Both these projects are based on a model of a high resolution, peripatetic digitization service to libraries using good conservation handling methods in which image capture equipment is transported to the manuscripts rather than the manuscripts to the equipment.

2.2 Internet Library of Early Journals

This is a joint project in the FIGIT Digitization programme involving Birmingham, Manchester, Leeds and the Bodleian Library in Oxford, with Oxford and Leeds as the lead sites. The initial goal is to provide Internet access to c. 140,000 page images of substantial runs (minimum 20 years) of three 18th century and three 19th century journals [3]. Two methods are being used for image capture. The Minolta PS300 book scanner is being used for converting original hardcopy, and a Meckle scanner to digitize microfilm with a view to employing this technique to digitizing our large stock of microfilm. Here, high resolution is less of an issue and we are scanning at 100-300 dpi and 256 shades of gray, and using JPEG compression so the images are comparatively small at 50-100 Kb.

Three indexing strategies are being assessed. Firstly, though we have had difficulties resulting from the poor quality of early type faces, we have scanned using OCR and indexed the full text, both with and without fuzzy-matching software. Secondly, we are linking in electronic indexes that already exist, viz. Periodicals Contents Index. Lastly, we are keyboarding electronic versions of the printed indexes and contents pages to enable browsing.

2.3 Broadside Ballads

Many works of poetry and music were published in earlier centuries in the form of Broadside Ballads, of which the Bodleian Library has a collection of some 30,000. HEFCE has funded a three year project to complete the creation of a machine-readable catalogue and to link the records to scanned images. The images have not been scanned in-house but given to an agency.

2.4 Electronic Readings in Management Studies (ERIMS)

This is a joint project in the FIGIT On-Demand Publishing Programme between Oxford, Sheffield and Aston Universities and publishers including Blackwell, OUP, Reed MultiMedia and Routledge. The objective is the production of Electronic Reading Packs in Management Studies as an alternative to the paper reading packs and short loan collections currently used by students. The corpus will eventually comprise 1,000 journal articles and book chapters; an estimated 20,000 images. User authorisation and copyright tracking software is being evaluated.

2.5 Beazley Archive Projects

The Beazley Archive[4] is an outstanding collection of some 5,000 photographs of Greek vases with associated text records owned by the Faculty of Literae Humaniores. These have been linked to an existing database of 55,000 text records which has been available to registered users for many years. More recently, project RAMA (Remote Access to Museum Artefacts), which is part of the European Union's DG XIII RACE Telecommunications Programme, has created 12,000 photographic images (300 dpi with gray scales) from the Beazley Archive.

A related project has also been undertaken in collaboration with British Telecom with the objective of creating a multimedia catalogue of exhibits in the Cast Gallery of the Ashmolean Museum by digitizing the existing paper catalogue together with video clips of exhibits. The technical emphasis has been on the potential for high bandwidth transmission of visual images and the project has entailed cabling the Cast Gallery for ATM.

2.6 Toyota Project

The Toyota car company has funded a project to create 8,000 colour images of material pertaining to cars from the Bodleian Library's John Johnson Collection of Ephemera. The technique used was to create 35 mm colour slides and then digitize them using an outside contractor rather than investing in our own equipment. Converted from Photo-CD to JPEG, the files are about 200 Kb and will display at 1536x1024 pixels. Originally intended as a standalone CD-ROM application, we have now developed this for WWW incorporating thumbnail images and SGML catalogue record.

Several points are worth making. Firstly, copyright is a key issue in the digital world and we must wait until it is resolved before we can estimate realistically the extent to which the move in library strategies from holdings to access provide significant cost benefits. Certainly, our experience has been that the provision of electronic publications to our users may have improved accessibility to information, but has not saved money. Indeed, to provide them at all we have had to invest heavily in IT, often stealing from our acquisitions budgets to do so.

President Clinton's Intellectual Property Working Group Green Paper has been of great interest, but in promoting the 'pay-per-use principle' it did not provide for the concept of 'fair-use'. Unsurprisingly, it was well received by the publishers. What is needed is to balance author/publishers' compensation with user/library interest in having access to information on fair and reasonable terms.

With respect the Oxford projects I have mentioned, only the ERIMS project is dealing with material where intellectual property is an issue. Here, the four publishers have agreed to provide copyright clearance and to enter into negotiations for a specimen license for the storage and use of these materials. Copyright for the other, older material rests with the holding institution. The position of the Bodleian Library is that its material should be made freely available on the Internet in accordance with its concern to provide for the international research community (and the requirements of FIGIT that the 'deliverables' should be freely available to the UK academic community), but it should appear with a suitable copyright notice and it will pursue infringements with vigour.

Secondly, file-store requirements for digitization are considerable and we have had to invest heavily in the necessary hardware. Fortunately, Oxford University has installed a Hierarchical File-Server with a current capacity of 13 terabytes and, theoretically, with no upper limit. Files are being 'staged' according to how frequently they are needed but it is expected that however deeply a file is archived, its delivery will commence within a minute from the request.

Thirdly, the world has moved to the WWW as the basic delivery mechanism for global information. Although we believe the WWW is the most appropriate system for the delivery of information at this point in time and are using it for the projects referred to above, we have reservations about its longevity and capabilities, as will be discussed below. We believe that there will be life after the WWW so our strategy has been to keep our data in formats that are not dependent on it in order that they should endure. We should advise others to do the same for information of a permanent nature.

3. CONNECTIVITY: THE DEMOCRATIC NETWORK AND CLIENT/SERVER COMPUTING

Speaking in general terms, there is now the expectation that not only should libraries be interconnecting, but also that the information services we provide locally and those remote services to which we subscribe should reach beyond library buildings to the users' desktops to wherever they are on campus. It has been our experience that, by and large (and choosing my words very carefully), suppliers of commercial information products, software developers and our users do not always fully appreciate the nature of modern, academic, networking environments. We are now in a world of distributed information which demands open computing systems. We have moved beyond the LAN, where the choice and configuration of workstations can be under the control of the service providers, to a 'democratic' environment of mixed networking protocols and heterogeneous workstation platforms. Unlike corporations and providers of national networks who may specify what protocols will be used on a network, most UK universities colleges and departments have developed their own LANs without sufficient consideration to wide area connectivity. In line with the excessively democratic nature of UK universities, which I suspect is true of Indian universities too, we find that their campus networks are equally democratic. In Oxford and elsewhere, there are many network protocols in use at the same time—e.g. TCP/IP, IPX, AppleTalk, NetBios, DEC Pathworks and NetBEUI—and a multiplicity of different workstation platforms.

Since connecting our Library to the University backbone in 1991, we have been grappling with these problems. There has been great demand to access our IT services from other libraries, departments, colleges and offices across campus. I shall illustrate the difficulties we have experienced by reference to CD-ROM networking.

One of our major concerns has been to provide as wide access as possible to our CD databases. We have had for some years a Novell/Opti-Net system which seemed marvelous at first but has proved to be limited in terms of universal connectivity. Firstly, maintenance of this system has proved to have high administrative overheads with respect to both installation of 'network-unfriendly' software and support of remote end-users. Secondly, it has been a source of frustration that we have

not been able easily to provide access to it from workstations other than PCs and this has disenfranchised many departments and colleges which have no PCs and are predominantly UNIX or Mac or are still based on mainframe VMS systems. Even for PCs it has been difficult to support remote users as there are certain configuration requirements that may conflict with their normal set-ups.

We are now in a world of distributed information and open systems which demands a client/server approach to computing. We sought a new delivery platform for our databases based on client/server architecture. In 1994, we chose SilverPlatter's Electronic Reference Library (ERL)[5,6] as a system that was both powerful and, in terms of accessibility, open, and with the flexibility to integrate it with our other information services. As far as we are aware, it is still the most advanced system for networking CDs. I should like to mention here the key features of ERL.

- ☐ Client/Server architecture which enables access from different workstation platforms using the standard network protocol, TCP.
- ☐ Client software for the major platforms provides a single user interface for the databases mounted on ERL, viz. PC-SPIRS, WinSPIRS, MacSPIRS, UNIX-SPIRS.
- ☐ Telnet access to the UNIXSPIRS client means that ERL can be used by virtually any user with a network connection, including our VMS mainframe community.
- ☐ The ability to search across different databases simultaneously. This has been of great benefit to the users, especially as much research now takes place at the boundaries of the traditional academic disciplines.
- ☐ Searching across databases mounted on different ERL servers located locally or anywhere on the Internet in a manner that is transparent to the user. This gives the service provider different options for sources of information which may be using different charging mechanisms. To illustrate, it would be possible for the user to search across Medline mounted locally for which the institution pays the subscription and to a top up file of the most current material held at the NLM for which charging is by connect time, and which charge could be passed to the end user. At the same time the user could be extending the search across Biological Abstracts provided as a BIDS-type national dataset held on a server located elsewhere in the UK.
- ☐ The ability to download CDs to magnetic disc. This is an attractive option now that magnetic storage is really no more expensive than CD drives and it provides marked increase in performance, especially for large databases with heavy usage.
- ☐ The possibility of mounting non-SilverPlatter databases on ERL. ERL uses the DXP (Data eXchange) protocol. Although other database suppliers have not adopted this protocol for server software, SilverPlatter has established the concept of ERL-compliance. This means that information providers can produce versions of their databases that can be mounted on an ERL server and accessed using the SPIRS suite of client software.

With the advent of the Internet, there has been the general acceptance that workstations must have TCP/IP support for wide area networking, and this has eased our problems with respect to transport protocols in a modern academic environment. As regards the workstations themselves, it is commonly asserted that in order to support the majority of departments in a University, information providers should cater for the three major platforms: IBM-type PCs, Macs and UNIX machines. We have found that, in reality, this is a simplification. PCs vary widely in their capabilities and might be running different operating systems ranging through DOS, Windows, Windows 95, Windows NT, OS/2 and even UNIX. In the UNIX world, we find that software will often have to be recompiled if it is to run on the major varieties of UNIX of which there is a bewildering variety in common use: SunOS, Solaris, HP, AIX, LINUX, UNIXWARE, SCO, HP and Digital UNIX, to mention a few. Even in the Mac world, whose ardent users often stress the ease of installing software, we can find machines ranging from those with 2-4 MB RAM and the System 6 operating system

and PowerPCs with 16 or more Mbs running System 7—we must often make a distinction between a 'Mac' and a 'Big Mac'.

Despite the attractions and technical advantages of the client/server model in providing wide area access to distributed information services, the distribution of new and updated client software to the users can be a nightmare. For ERL alone we are currently maintaining nine different versions of the SPIRS and these do not meet the needs of all our users. We must also provide a wide array of client software for other systems, such as Lexis/Nexis, CrossFire for Beilstein, and Acrobat readers for PDF files—a format that is of increasing importance in the world of electronic publishing. Each user or local network administrator must obtain them all to have access to the full range of our services. The implications in terms of end-user support are obvious and we often refer to this as 'the down-side of client/server computing'.

4. INTEGRATION OF NETWORK RESOURCES AND THE WORLD-WIDE WEB

The WWW has brought a hitherto unimaginable degree of integration to the Internet, and we, as has much of the Internet, have adopted it as a principal delivery mechanism, for searching and navigation and for integrating our information services. It has even been possible for us, through the development of WebSPIRS to provide WWW access to our ERL databases, so ameliorating the problems of distributing client software. This is an example of a WWW gateway. It is an ERL client that runs on a WWW server using the CGI standard interface and anyone with a WWW browser can access the WWW server on which it has been mounted which in turn is able to interrogate the ERL server.

Two important features have contributed to the widespread adoption of WWW. Firstly, in itself, the WWW is just a browsing tool for distributed information, but the CGI (Common Gateway Interface) standard specification enables its functionality to be increased dramatically so that the WWW server can run scripts which enable it to interface with other information systems such as database search engines mounted locally or on remote hosts.

Secondly, the development of the graphical browser, Mosaic, at NCSA, with its abilities to handle HTML documents, images, audio files and other digital formats, and to access information by a wide variety of the TCP protocols, e.g. ftp, gopher, news, http, telnet, etc. Browsers like Netscape and Internet Explorer are continuing to add functionality such as a mail client, support for JAVA, ActiveX and the ability to view Acrobat files without having to install a separate Acrobat reader. WWW browsers are coming closer to providing a complete desktop environment able to access everything.

However, our experience has shown that there are serious limitations to the WWW. Firstly, the HTML format for Web document is really a simple mark-up language focused on presentation. The HTML standard is continuously being extended so it can, for example, handle tables and wrap text around images. But in terms of semantic content, it is extremely weak with tags just for <TITLE> and <ADDRESS>, and this seriously limits its usefulness for structuring documents.

Where we have structured data for which we wish to provide a sophisticated search interface, for example files in SGML or a database format, our strategy is to hold the data in whatever format we judge most suitable, and to use the CGI interface to extract the data and convert it on the fly to HTML files for delivery. This is the technique employed by WebSPIRS and we are using it to provide WWW access to our image and other databases mounted on the freeWAIS-sf, PAT, SQL and Allegro platforms [7].

The WWW's http protocol 'stateless', that is, it serves files on request and retains no record of transactions by a particular user that would enable it, for example, to build a complex search. From the viewpoint of the information professional, it cannot compete with 'session-based' protocols (often now referred to as 'stateful') like ERL's DXP or Z39.50 which are designed to interrogate bibliographic catalogues and indexes. As we have seen, WWW

Gateways can compensate, but this technique is inherently flawed in that they entail overheads in terms of systems development and maintenance, and almost inevitably there is a loss of functionality. Also, they require additional processing power and, potentially, they make inefficient use of the network.

We foresee the development of 'multi-protocol servers' which are able to process requests and deliver results in the format appropriate to the requesting protocol directly to the client rather than passing through a WWW gateway. A working example is the GN server that can handle both Gopher and http requests [8]. In this way, it should be possible to maintain a single version of the data. Further, such multi-protocol servers should be 'extensible' in that it should be possible to load support for new protocols when they are developed. It should be possible to do this dynamically, without interrupting the service.

There has also been a general perception of a need for extensibility of client software, i.e. the browsers. NCSA has worked on CCI, the Common Client Interface for Mosaic, intended to provide a standard interface by which all WWW browsers can communicate with external programmes. Similarly, there is a proposal for World-Wide Web Applets (W3A) that would be a standard API (Applications Program Interface) that would enable software to be sent by the server and run by the WWW browser[9]. However, in this area, proprietary solutions that looks set to become a de facto standards are Sun's JAVA and Microsoft's ActiveX. They effectively enable programs to be downloaded to clients so extending their functionality dynamically. For example, they enable the Web to be used to deliver specialist software such as ERL or Z39.50 clients to the user extremely easily, and also viewers for documents held in any format, for example, SGML and common wordprocessor formats such as Word or TeX. We expect a future scenario in which browsers will be able truly to provide a single interface for all their information requirements. They are increasing enormously the potential of electronic publishing and we shall need to redefine our models of what a learned publication comprises.

5. CATALOGUING AND INDEXING INTERNET OBJECTS: UNIFORM RESOURCE IDENTIFIERS AND META-DATA

Turning our attention now to Internet standards and the problem of cataloguing and indexing distributed Internet sources, a rough division can be made between methods of manual compilation and those on automated techniques[10]. There are many fine examples of the former, some using traditional classification schemes such as Dewey and LCSH. The main problems are generally scope and sustainability. Often they start from the efforts of a single individual but in order for them to be maintained, it is usually necessary for collaborative effort and for resources to be found if they are to be extended beyond small special collections.

In contrast, there is a range of mechanisms currently under development using automated means to improve information retrieval on the Internet. As a group, they are often referred to Robots, Spiders, Wanderers, Web Crawlers and Worms. They employ differing strategies in gathering the information to index: they will follow links between documents in a given range, extracting information, typically the URLs, titles and some or all of the content, and feed it to a database for indexing.

A good compilation and description of the different approaches of known robots is held by Martin Koster at Nexor[11]. Many of the robots now available are listed on the CUI's page on W3 Search engines[12]. The better known include NASA's RBSE, the Internet Research Task Force Research Group on Resource Discovery's Harvest, Lycos, Aliweb, Alta Vista, Web Crawler and the World-Wide Web Worm.

Despite what many people think, robots do not actually traverse the Internet following links and collecting information to index, but they stay at home and bring across the network all the information that will be indexed. Robotic systems like Lycos can potentially impact

network performance and place unreasonable loads on servers. Network etiquette now demands that they should respect guidelines that have been developed in this regard[13]. It is generally thought that in their present form they will not be sufficiently scalable to provide global indexes.

In addition, their search patterns follow simple rules, say, to follow a set number of links outward from starting URLs, which means they are undiscriminating in their selection of information and they may index documents whose authors would not have wanted them indexed or which are not appropriate for indexing, e.g. ephemeral material such as 'messages of the day'.

Robots are undoubtedly powerful tools but limited in their ability to provide sophisticated indexes. To a great extent this is due to the information they are indexing, which may or may not be organised, and is of variable scope, quality and structure—they represent 'dirty searching'. Techniques, such as the use of expert systems, to improve their performance will continue to develop, but at present, it appears that large-scale indexing of distributed sources requires a mix of robotic means and direct human intervention by relevance feedback, or, (as with ALIWEB), in the preparation of the source data used to compile a global index.

The Internet Engineering Task Force (IETF) is the body in which Internet standards develop. One of its sub-groups, the IETF-URI, has been working on draft for standards concerned with the retrievability of Internet objects. URI stands for Uniform Resource Identifiers, and though nomenclature has varied, this term is now used generically to include, URLs, URNs, URCs and URAs[14].

Briefly, we are familiar with URLs, Uniform Resource Locators, which describe the location of an object in the Internet and the protocol used to access it. In order to make retrieval more efficient, the concept of an URN (Uniform Resource Name) is being developed. It is proposed that schemes are created whereby objects would be registered with a name server in a system of name servers, and issued with a unique identity. It is the URN that would be specified when a resource was requested, which would then be resolved into an URL by a name server.

An object on the Internet may be available at many locations therefore it would be possible for the name server to provide the most appropriate URL, for example, the 'closest' in network terms, thereby optimizing use of the network. Assuming the name servers were notified, so it would be possible for items still to be found when their locations were changed. URNs can be seen as analogous to ISBNs and it is likely that only items considered to have a certain value would be registered.

OCLC has developed a cataloguing standard that would appear to be a first step towards implementing URNs. Called the PURL (Persistent URL) it is functionally an URL, but instead of pointing directly to the location of an Internet resource, it points to an intermediate resolution service which associates the PURL with an URL and completes the appropriate network transfer. Thus a PURL can stay constant and only the resolver need be notified when an URL changes.

URCs, Uniform Resource Characteristics, provide meta-data or meta-information about an object, and is analogous to bibliographic records. There has been discussion of what information should be provided in a URC, which took a step forward at a meeting arranged to discuss this specifically: the OCLC/NCSA Metadata Workshop Dublin, Ohio, March 1-3, 1995[15]. It proposed a core set of elements to appear in a URC. Containing the elements, Subject, Title, Author, Publisher, Other Agent, Date, Identifier, Object-type, Form, Relation, Language, Edition, Source and Coverage, no element is obligatory and all can be repeated. A second workshop was held in Warwick, UK, in February 1996.

The group took into account related work that had been done by others, notably, that resulting from the TEI (Text encoding Initiative) which has made proposals for mark-up of meta-data in SGML documents in the area known as the TEI Header[16]. In the world of library science, the 856 field was added in

February 1995 to the USMARC format especially to hold information pertaining to electronic publications and both the Library of Congress and OCLC[17] have been using it for cataloguing.

We must wait to see how these standards develop and the extent and manner in which they are adopted before we know if they can serve as a basis for wide-scale indexing of the Internet. There are many further issues that will need to be resolved in order for this to be achieved, such as whether the production of the meta-data can be automated, the extent to which authors will be empowered to provide their own meta-data, and the place of controlled languages in such schemes.

6. REFERENCES AND NOTES

1. Follet report. Joint Funding Council's Libraries Review Group, 1993.
2. Law, Derek. Staying at the leading edge. *Library Association Record*, 1995, **97**(12).
3. Blackwoods Magazine, Annual Register, The Builder, Notes and Quries, Gentleman's Magazine, Philosophical Transactions.
4. http://www.ashmol.ox.ac.uk/castgal/cghome.html
5. http://www.silverplatter.com/
6. Price, David J. ERL on campus, December 1995, 101, 29-36.
7. Examples can be found at http://www.bodley.ox.ac.uk/cgi-bin/rspnew.tcl?search (Refugee Studies Programme database) and http://www.bodley.ox.ac.uk/cgi-bin/toyota.tcl?search (Toyota Imaging Project)
8. http://hopf.math.nwu.edu:70/
9. For information on the CCI specification, see http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/cci-spec.html; for W3A, see http://www.let.rug.nl/~bert/W3A/W3A
10. Price, David J. Indexing the world : Current developments in accessing distributed information. *In* Wissen in elektronische Netzwerken. Proceedings of the 60th Annual Conference of the German Classification Society. HChr Hobohm and HJ Watzen (Eds), Oldenburg, 1995.
11. http://web.nexor.co.uk/mak/doc/robots/active.html
12. http://cui.unige.ch/meta-index.html
13. http://web.nexor.co.uk/mak/doc/robots/guidelines.html
14. http://www.ietf.cnri.reston.va.us/ids.uri.html.
15. http://www.oclc.org:5046/conferences/metadata/metadata.html
16. Sperberg-McQueen, CM and Burnard, Lou (Eds). Guidelines for elctronic text encoding and interchange (TEI P3). Chicago and Oxford, ACH-ALLC-ACL Text Encoding Initiative, May 1995. (2V).
17. Interestingly, OCLC have designed a WWW tool called Spectrum to allow individuals with no specialized knowledge of cataloguing to create bibliographic records of electronic resources which will convert with reasonable success between USMARC, TEI Headers and URCs. See D Vizine-Goetz *et al*, 'Spectrum : a Web-based tool for describing electronic resources'. *Computer Networks and ISDN Systems*, 1995, **27**, 985-1001.