Content and services issues for digital libraries

By Derek Law

Introduction

Most of the attention in digital libraries has gone on the technical developments which do and will make them feasible, whether permanent locators, version control, metadata or cross platform searching. Thus far much less attention has been given to the areas of e-collection building and on-line services. What little thinking has gone on is limited in vision, can be partial and even part of a lopsided political agenda (Sun, 2003). A recent review of the history of "Informatization" over the last forty years gives collections barely a mention (Duff, 2003). But the sheer volume of electronic materials is growing rapidly (OCLC, 2003) and requires thought to be given to policy on collection building as well as the technology and practices which will allow it to happen. Present academic research builds on the collections of the past: it then behoves us to build collections for the future. It has been claimed in the context of Open Archives that "the biggest challenge is getting content" (Pinfield, 2003). In the context of e-collection building the challenge is perhaps that of building collections of coherent content.

Building research collections for the future

In the past building collections was relatively straightforward. The papers of distinguished academics were collected from their studies after death; manuscripts and books were purchased from rare book and manuscript dealers, contacts were cultivated in the hope of donations. The vary stability of the paper record allowed patience and often multiple opportunities to determine where papers gravitated to. Nor were the collections only paper, but sometimes physical objects. The University of Hull famously added Philip Larkin's lawnmower to its library collections (Guardian, 2002).

The issues are much more daunting when it comes to electronic materials and largely revolve around media formats and preservation as described in the chapter by Lazinger. But we do precious little in terms of what would have constituted collections in the past. E-drafts of documents and paper; e-correspondence between researchers; personal files on a pc rather than in a filing cabinet; the electronic equivalent of lab books are all falling through the net. While we feed off the collections of the past we generally fail to reflect on how the born digital collections of the future will look. Nor do we consider how the material will be held. The absence of agreed repository standards must be a major cause of concern. Ironically, as in so many things one can see a potential solution in looking back to the experience of the past to develop thinking on the future. The Maori tradition is an oral one and they have developed a quite specific set of criteria to guide the selection of the keepers of that oral tradition:(Winiata, 2002)

1. Receive the information with utmost accuracy
2. Store the information with integrity beyond doubt
3. Retrieve the information without amendment
4. Apply appropriate judgement in the use of the information

5. Pass the information on appropriately

These seem a perfect guide to the preservation requirements of tomorrow's e-collections.

A typology of collections

Thinking on collections has most fruitfully taken place within the context of the Digital Library Federation (http://www.diglib.org/dlfhomepage.htm) and they have produced interesting work for example on strategies for developing sustainable and scalable digital library collections. Greenstein (2000) proposes four types of collection

- local digitization projects that produce surrogates for analogue information objects;
- data creation projects that produce information resources that have no analogue equivalent and are in this respect "born digital";
- the selection of existing third-party data resources for inclusion in a collection either through their outright acquisition or by acquiring access under some licensing arrangement; and
- the development of Internet gateways comprising locally maintained pages or databases of web-links to third-party networked information

This typology allows an exploration of the nature and extent of what is, should be and could be made available.

Digitised Surrogate Resources

It is a commonplace that not all existing collections will be digitised. Scale, copyright and value are argued to make such conversion implausible. It is certainly the case that at present we tend to see projects delivering selected subsets of collections rather than the whole. Digitised resources can be further sub-categorised beyond Greenstein's single overarching category, because the motives for digitisation are very varied. Improved access, preservation, aggregation of scattered material, and more are all reasons for creating digital collections, as the following examples of the sub-categories show.

Surrogates of rare items: the British Library

An excellent example of this is the British Library's Treasures collection (http://www.bl.uk/collections/treasures.html), where rare treasures are made more accessible to the public (and indeed to scholars). This collection contains such heterogeneous material as Magna Carta, the Lindisfarne Gospels, the Gutenberg Bible and the notebooks of Leonardo Da Vinci. What these great documents have in common is their rarity and their public prominence. The e-collection acts as a surrogate to allow these great iconic treasures to be open to all.

Surrogates for whole or part collections: The Springburn Virtual Library

During the summer of 2000 it became apparent that the Springburn Community Museum faced closure for financial reasons. Although the collections were to be transferred to the Mitchell Library in Glasgow, this much loved and popular local resource would be separated from its community. A project was put in place to ensure

that the museum's rich collection of local photographs would still be accessible to the local public over the internet. Funding was secured to digitise a representative selection of materials from the collections and to lay the foundations for Springburn Virtual Museum. Images were chosen to convey the social and economic history of Springburn, notably community and tenement life and the important local railway industry. http://gdl.cdlr.strath.ac.uk/springburn/. As a result a community threatened with the loss of a resource has had at least a subset of it made more accessible to all.

Digitised surrogate collections assembled from multiple repositories: the Valley of the Shadow

The much admired Valley of the Shadow Project focuses in great detail on the experience of two communities, one Northern and one Southern, through the American Civil War, as an exemplar to give an understanding of the experience of the nation as a whole. It consists of a hypermedia archive of sources for Augusta County, Virginia, and Franklin County, Pennsylvania. A rich variety of materials has been assembled - newspapers, letters, diaries, photographs, maps, church records, population census, agricultural census, and military records. It encourages users to interact with materials rather than simply access them. http://www.iath.virginia.edu/vshadow2/

A collection with a quite different focus and ambition is the Great Britain Historical GIS Project (www.gbhgis.org), which aims to have systematic information on the history of every locality in Britain, using everything from Ordnance Survey maps to Victorian gazetteers and Defoe's *A Journey through the Whole Island of Britain*. It can be searched using postcodes and aims to allow everyone to access information relevant to their own area.

Collections assembled specifically to be digitised

The Aspect project (http://gdl.cdlr.strath.ac.uk/aspect/) was set up to create a digital archive of the ephemera - leaflets, flyers, postcards, newsletters - produced by candidates and political parties for the first Scottish parliamentary election in May 1999. The archive is based on the collection of election ephemera held by the Andersonian Library at the University of Strathclyde, which is acknowledged to be an important and unique record of a key event in Scottish history. The creation of a digital archive will significantly improve the accessibility and usability of the information contained within the collection whilst conserving the original materials, which may be subject to deterioration through loss and damage. Thus a collection being built for use by future researchers is being made immediately available, using digitisation as a deliberate strategy in acquisition.

Born Digital Resources

The number and scale of these is growing from scholarly journals to new fiction, from datasets and satellite images to digital video and computer generated graphics. Many are being preserved. But examples of born digital collections are rare. It is arguable that these remain individual items rather than forming a coherently built collection. Perhaps the nearest to this is the various collections of learning objects being assembled in many universities. For example, Boezerooy (2003), gives a comprehensive overview of the Australian experience which demonstrates that these exist but are not always created with library advice or assistance or indeed even with long term preservation in mind.

Third Party Data Sources

In the UK, JISC began its work of building the Distributed National Electronic Resource in 1990 (Law, 1994) and now has a hugely rich collection of resources licensed to the community (JISC,2003). That consortial licensing model has been widely followed. The International Coalition of Library Consortia (ICOLC) first met in 1997 and has grown to be a self-help group of some 150 consortia from all over the world, which considers issues of common concern, principally in the context of higher education and research. Without necessarily supporting it however, ICOLC (http://www.library.yale.edu/consortia/) in effect works within the present pattern of scholarly communication to make material as available as possible.

The electronic environment offers up new and as yet unexplored models of data acquisition, whether for a single institution or in consortia. The intention expressed by Singapore in its seminal planning for the Intelligent Island (Chun Wei Choo, 1997) is to create an information entrepôt and hub for the region. It is easy to build on this concept to develop the concept of information arbitrage (Law, 2001), the notion of buying and selling information around the world, taking advantage of time shift to buy data cheaply at offpeak times when little used in a country. Similar thinking has informed the development of 7x24 reference services as described later in this chapter.

Quite novel models have also been proposed to allow freer access to the scholarly research literature. Most of the debate has centred on the ailing STM model more fully explored by Harnad in his chapter The model he has advocated for many years has moved from the fringe of debate to the mainstream. Most recently the so-called Budapest Declaration, under the aegis of the Soros Foundation declared that:

> "We invite governments, universities, libraries, journal editors, publishers, foundations, learned societies, professional associations, and individual scholars who share our vision to join us in the task of removing the barriers to open access and building a future in which research and education in every part of the world are that much more free to flourish." http://www.soros.org/openaccess/read.shtml

Most of the debate has focussed on the perceived failure of the STM (Scientific Technical and Medical) model of scholarly communication where the highest priced journals exist. Many other initiatives such as Biomed Central (http://www.biomedcentral.com/) and SPARC (http://www.arl.org/sparc/) have demonstrated the concern felt in the wider scholarly community at the present state of scholarly communication and the need to change that. We appear to have developed a monster which has steadily lost sight of the fact that publishing exists to support research and not the opposite. But in this debate little thought has been given to the Humanities and Social Sciences, where huge numbers of journals and researchers exist and where journals are often effectively produced as a labour of love from within university departments. Here some steps are being taken actively to persuade and assist small scholarly publishers to shift their content to electronic formats. The role then is to mediate the transfer to an e-environment and not simply to acquire content. Such an initiative is the SAPIENS (Scottish Academic Periodicals: Implementing an Effective

Networked Service) project involving six Scottish universities and the National Library of Scotland (http://sapiens.cdlr.strath.ac.uk/).It aims to:

- examine the case for a centralised Scottish electronic journal service which might enable and encourage smaller publishers to make existing and new journals available in electronic form

- design and build a demonstrator service, which will deliver current journals from a representative selection of publishers via a common gateway

- develop and launch an operational service, together with a marketing strategy to ensure that it is self-sustaining within a year of the end of the project.

Librarians here as elsewhere have developed a catalytic role in helping to make available the content required by library users.

Mirroring and caching

This is a somewhat neglected subset of third party licensing. An excellent early example of this is the Visible Human dataset. This was originally constructed in the United States with the support of the National Library of Medicine (NLM). It contains images of a 39-year old convicted murderer who, prior to his execution, donated his corpse to medical science. The dataset was subsequently expanded with the addition of the images of a female at greater resolution than for the male. The bodies have been "sliced" to create the images. NLM did not want to see copies of the dataset mounted outside the USA, quite properly fearing that issues such as version control and quality assurance were not sufficiently settled in the mid-1990s to give comfort of proper data management. For the UK this proved a problem since this wonderful resource was heavily used in medical teaching and consumed great quantities of bandwidth as images were slowly downloaded. Mirroring was the obvious solution. Discussions began with NLM and after protracted discussions the final sticking point (according to folklore!) was the need for guarantees on what would happen to the data if the host institution disappeared. At that point, in 1997, JISC accepted an offer from the University of Glasgow to act as the host (http://vhp.gla.ac.uk/), not least on the grounds that it had already existed for half a century before Columbus sailed the ocean blue. Whether or not the tale is true, it does demonstrate that mirroring can be just as complicated an exercise as licensing commercial data. Certainly in the UK, as network charging begins to influence decisions, it seems reasonable to expect a greater interest in mirroring s a method of reducing traffic as much as improving accessibility.

The same is true of caching data. This is one of the black arts of computing but does have a significant impact on costs, traffic and availability. The UK National Cache has been studied in depth (Sparks et al, 1999) in terms of performance and value for money and this is very informative in indicating the impact that an institutional caching strategy might have.

Internet Gateways

Such gateways have now existed for several years, whether as generalist services such as BUBL "Free User-Friendly Access to Selected Internet Resources Covering all Subject Areas, with a Special Focus on Library and Information Science" (http://bubl.ac.uk/) or subject specific services such as EEVL for the engineering community (http://www.eevl.ac.uk/). Typically these are university based "free" services, funded by third parties, often government agencies. These are based on the notion that no single institution can manage with discrimination all the information on the Internet and that the labour can sensibly be divided. The UK experience began with several projects under the access to networked resources strand of the Follett Report (Law & Dempsey, 2000). These were intended to cover a range of subject areas: OMNI (medical and bioscience), ADAM (art and design), EEVL (engineering) and RUDI (urban design), all began the task of building databases of Internet resources in their respective subject areas from scratch, while SOSIG extended a pre-existing project. Funding was also provided to support the gateways by funding ROADS, which aimed to develop software that could be used by the gateways to create the resource databases and serve them to users via the Web. The success of these initial projects led the JISC to develop the Resource Discovery Network (RDN), which uses this approach to cover all subject disciplines (Dempsey, 2000). The usage of the RDN gateways has been disappointingly low and this national approach may have to be reappraised.

One major issue appears not to have been addressed so far. There is a bland assumption that there is an almost infinite supply of bandwidth and that issues of access and slow to load pages will disappear: that view is not necessarily shared by all. At the same time there is an equally unthinking assumption that resources are either good or bad. But there is a more sophisticated but so far neglected approach which asks whether the Pareto Principle (the 80:20 rule) might also apply to on-line resources. It is typically assumed that access should be given to the best or most complete or most authoritative material, but these terms are never explored or defined. Networked environments add the complication of accessibility in a quite novel way. For example, in many parts of Europe, the quality of connectivity to the United States drops dramatically after the golden hours of the European morning, once American users wake up and begin to log on. So is a similar or smaller resource (but just as accurate) available twenty-four hours a day to be preferred to a larger resource effectively available for, say, only two-thirds of the day? There is a need for a much more sophisticated appraisal of all the factors surrounding internet gateway access than has perhaps been the case thus far.

Shared services

Internet gateways are perhaps closer to services than collections, although they will undoubtedly help to define the perception of the library in future. If libraries can provide online services, which are seen as independent, authoritative and right, they seem certain to see off competition from those less skilled. In an inversion of Gresham's Law, Law's First Law[1] states that "Good Information Systems will drive out bad". The development of electronic services in library dates back to the creation of the first automated systems in the 1960's and the area of e-services is well understood and much discussed, for example by Pantry & Griffiths (2002). Thinking is only just beginning on how using the network to share services can be exploited – although

interlending and document supply is a longstanding triumph of professional co-operation much enhanced by new technologies, as is shared cataloguing.

The development of shared programmes for information skills training is perhaps an old-fashioned but important starting point for sharing. A growing number of locally prepared but networked based products is available.

Much interest has been shown in shared reference services where timeshift allows 7x24 coverage for those staff and students who prefer anti-social habits to the normal working day. For example the University of Technology Sydney and the University of Strathclyde in Glasgow are piloting such a shared service where each answers reference enquiries from the others user's during the questioners night – daytime in the other country.

Conclusion

To some extent the issue of e-collections will define the future of libraries. At one extreme there is Brewster Kahle who has adopted the universal library philosophy of the great nineteenth century libraries considering the internet to be the library and with a very unsentimental view of past glories such as the Alexandrine Library: "Great library – too bad it was burnt" (Kaushik, 2003). Less comprehensive virtual libraries will require the application of the traditional skills of selection of content as well as its preservation if not physical space, while the argument for the library as physical place even in a digital future has been strongly argued by the UK's Library and Information Commission (Library and Information Commission, 1999). Whatever the future holds for libraries in terms of physical location, e-collections will need to be built. It is then our existing professional skills in selection, acquisition and cataloguing which place librarians as the best qualified group to organise content – provided the challenge is recognised and accepted.

Notes

1. The creation of Law's First Law is as much an attempt to seek attention as succinctness. There is also Law's Second Law, which emphasises the importance of offering information skills training through the Library. It states that "User friendly systems aren't"

References

Boezerooy, Petra (2003) Keeping up with our neighbours: ICT developments in Australian Higher Education LTSN Generic Centre, [n.p.]

Chun Wei Choo (1997) IT2000: Singapore's Vision of an Intelligent Island. Book chapter in Intelligent Environments, edited by Peter Droege Amsterdam, North-Holland

Dempsey, Lorcan (2000). The subject gateways: experiences and issues based on the emergence of the Resource Discovery Network. Online Information Review, 24(1), 2000. p. 8-23. Also available at http://www.rdn.ac.uk/publications/ior-2000-02-dempsey/

Duff, Alistair S. (2003) Four "e"pochs: the story of informatization. *Library Review* 52(2) pp58-64

Greenstein, Dan (2000) Strategies for developing sustainable and scalable digital library collections. http://www.diglib.org/collections/collstrat.htm

Guardian (2002) *The Education Guardian* Thursday May 9, 2002 [news item] http://education.guardian.co.uk/higher/humanities/story/0,9850,712877,00.html

JISC (2003) e-collections: exploiting the opportunities [JISC collections folder] Bristol, JISC, 2003 www.jisc.ac.uk/collections/

Kaushik, Radhika (2003) Spreading The Digital Word *ExtremeTech* 29 April 2003. http://www.extremetech.com/article2/0,3973,1047454,00.asp

Law, D. (1994) The development of a national policy for dataset provision in the UK: a historical perspective. *Journal of Information Networking* 1(2) pp103-116

Law, D. (2001) The Library in the Market: information arbitrage as the new face of an old service. *IATUL Proceedings* Vol 11 (New Series) 2001. Delft: Delft University of Technology, 2002.

Law, Derek & Dempsey, Lorcan (2000) A Policy Context – e-Lib and the emergence of subject gateways *Ariadne* (25) 5pp. http://www.ariadne.ac.uk/issue25/subject-gateways

Library and Information Commission (1999) 2020 Vision http://www.lic.gov.uk/publications/policyreports/2020.pdf

OCLC (2003) Five year information format trends. OCLC, March 2003 www.oclc.org/info/trends/

Pantry, Sheila & Griffiths, Peter.(2003) Creating a successful e-information service. London: Facet

Pinfield, Stephen (2003) Open Archives and UK Institutions *D-Lib Magazine* 9(3), 2003

Sparks, Michael, Neisser, George & Hanby, Richard (1999) An Initial Statistical Analysis of the Performance of the UKNational JANETCache http://wwwcache.ja.net/papers/initial_analysis/

Sun (2003) Sun Microsystems Educational Consultation Forum. Creating the Distributed National Research Library, paper ECF04 February, 2003

Winiata, Whatarangi (2002) Ka purea e ngā a hau a Tāwhirimātea: Ngā Wharepukapuka o Ngā Tau Ruamano. Keynote address, LIANZA Conference, Wellington, 2002. http://www.confer.co.nz/lianza2002/PDFS/Whatarangi%20Winiata.pdf