Challenges for the Digital Libraries and Standards to Solve them

Alan Hopkinson

Abstract

There are numerous challenges in organising a digital library (DL) and successfully get to the user the articles he/she wants. Some of the problems have already been addressed and a few more are yet to be solved. Various standards have already been developed in storage and retrieval of digital data which are described here. They range from the standards that cover Portable Document Format files through the standards that govern international cataloguing efforts to standards for searching. Others are still under development and are described in the state in which they are currently. In conclusion, standards are necessary for every aspect of the digital library. New standards are being developed by the formal international and national standards bodies and one set up especially for the purpose such as COUNTER has been highlighted.

Keywords: Digital Library, Standards, Digital Storage and Retrieval, Digital Preservation, COUNTER

1. Digital Libraries and the internet

Digital libraries, has been referred to in this paper 'as any collection of digitised material'. Internet began to be used as a source of information and much informal material that could never have been published before became available. In industrial nations the universities were the first to be connected to the internet but by the year 2000 many homes in the developed world had internet access and the internet was being regarded as a source of information which was free of charge, at least free after having paid a fee to the telephone company for the connection. Eventually journals began to be published on a server accessible from the internet and instead of finding a journal in the library, potential readers could find and read articles on the internet. The publishers of the journals had sold journal issues to take into account the cost of production and the raw materials such as paper and ink. Purchasers realised that they could not have a printed journal for free. But because the internet was normally free, the end users did not understand why they had to pay for journal articles on the internet. However the publishers continued to charge the same price as they had for the hard copies. Because material is universally available on the internet the readers of journal articles thought that their libraries could provide everything for free. It is in this scenario that we find ourselves today.

The UK's Society for College National and University Librarians (SCONUL) has just published its annual library statistics for 2006-7 showing the proportion of digital to print journals shifting from 25% to 75% over the previous eight years. 45% of the acquisitions budget in UK Higher Education goes on electronic materials. (1). Librarians need more information than these bare statistics to be able to challenge the publishers. In some instances challenges have come through the open access movement under which articles that have been

published are mounted on servers accessible to the world, using software such as Eprints or DSpace.

There are several standards released and under development which make it easier to access digital libraries on the internet, retrieving and accessing the relevant articles. There are, for example, the internet protocols, PDF (2), library standards and conventions that govern the creation of bibliographic records such as Anglo-American Cataloguing Rules (3) and ISO 2709 (4) which specifies the structure of the MARC (5) record, the format in which catalogue records are exchanged between library automation systems and databases of references and it is also the format of the UNISIST Reference Manual (6) and the UNESCO Common Communication Format (7), which have been used to exchange references between libraries and secondary services databases.

2. Current Successes and Challenges

The PDF format is standardised as a proprietary standard and has been adopted as ISO 32000-1:2008 (8) Document management — Portable document format — Part 1: PDF 1.7. The standard itself states that it specifies a digital form for representing electronic documents to enable users to exchange and view electronic documents independent of the environment in which they were created or the environment in which they are viewed or printed. It is intended for the developer of software that creates PDF files (conforming writers), software that reads existing PDF files and interprets their contents for display and interaction (conforming readers) and PDF products that read and/or write PDF files for a variety of other purposes (conforming products).

ISO 32000-1:2008 does not specify the following:

- specific processes for converting paper or electronic documents to the PDF format;
- specific technical design, user interface or implementation or operational details of rendering;
- specific physical methods of storing these documents such as media and storage conditions;
- methods for validating the conformance of PDF files or readers;
- required computer hardware and/or operating system.

This is fine for normal digital libraries. Commercial libraries need mechanisms to control usage. Recent versions of Acrobat are more sophisticated alongside the PDF structure and can hold documents which cannot be printed or saved preventing unauthorised copying, the needs of publishers and authors being in mind.

One of the most important tasks for libraries providing access to on-line resources is the justification of their spending. This is especially necessary when the resource is available from more than one source. Not that all databases can provide usage statistics. But the statistics come in many different formats and with different categorizations. There is now a standard called COUNTER (9) which suppliers should adhere to. This standard will be discussed in greater depth later in this paper.

Searching systems is one of the most difficult operations to achieve satisfactorily. But searching has always been fraught with imprecision and always will be by its very nature. In the past there have been standards such as the Common Command Language where codes were given to particular activities and entities in searching. ANSI/ NISO Z39.58 - Common Command Language for Online Interactive Information Retrieval (10) released in 1992 but now withdrawn states it is "Useful to systems designers who want to specify a

uniform command terminology, describes nineteen non-proprietary command terms for use in online information retrieval systems." The equivalent ISO standard ISO 8777:1993, Information and documentation — Commands for interactive text searching[1] "specifies a basic set of commands for the interactive search of retrieval systems data and the types of response expected from the processing system is intended for use by designers and users of information retrieval systems, including computer-based library catalogues and computer-based database access and search facilities." However this kind of "command language" had the aim of standardising operations (it includes the Boolean operators AND, OR and NOT and proximity operators as well). In the days when development of online searching began there was a feeling that space should always be saved and codes should be used. A by-product of the codes is that they are not linked to any language. Nowadays user friendliness is the priority and this standard has been withdrawn.

Systems have become more complicated and users would often like to search across more than one database or collection of journals. For some time now there has been a standard which enables searching across bibliographic databases. It works reasonably well with library catalogue databases which are highly structured and use the MARC format[2]. The MARC format is one of the success stories of standards in the library world. It is not actually a standard but is under the ownership of a committee, MARBI, which is very closely associated with the US Library of Congress. However from the late 1960s the structure of the records though not the content has been the subject of a standard, ISO 2709 (11) (originally NISO Z39.2, Format for Information Interchange). It makes the exchange of catalogue records in machine-readable form quite standardised. Many secondary services do not use the MARC standard so there is less consistency in their format (for display and manipulation). Z39.2 and MARC provide the framework for records to be transferred between systems and then read by the end user through common software. This uses an antiquated record structure which dates from the days of tape transfers. What is not there is the facility for cross-system searching. That is provided by Z39.50 (ISO 23950 Information and documentation — Information retrieval (Z39.50) (12) — Application service definition and protocol specification). This is an attempt to standardise indexes and enable library automation system vendors to write an interface to their systems which any other system can access. When it comes to digital libraries as opposed to library catalogues, we are more interested in accessing records of articles and searching through full text. Cross system searching does not work well though companies have developed their own interfaces which are heavy on intellectual work to make an appropriate interface to data in many different formats. No standard are being developed in this area though NISO has a task group on Metasearch Initiative with subcommittees on Access Management, Collection Description and Search/Retrieve. In practice companies like Webfeat and MUSE have developed interfaces to all the different databases on an individual basis.

3. Standards in Progress

There are many bodies developing standards and norms and producing guidelines and conventions. Firstly there are the national standards bodies which are members of the International Organization for Standardization, ISO, for example, British Standards Institution, National Information Standards Organization (NISO) in the USA and

Bureau of Indian Standards here. However, these bodies react on the requirements of professionals, librarians or publishers who are members of their committees. Currently, NISO is the most active body and it has on its committees librarians from US libraries across the different library sectors as well as a few foreign members. Many standards they develop feed into ISO work sometimes being purely adopted as an ISO standard

NISO has a large number of different interests in its standardization activities; its website lists

- library technical services;
- the acquisition and management of e-resources;
- library systems implementation including ILS, ERMS, link resolvers, and web interfaces;
- cooperative electronic arrangements with other libraries, consortia, or content providers; or
- long-term preservation activities.

3.1 Usage

As far as the management of e-resources is concerned the library needs to know how much usage there is in order to justify the continued purchase of licences to the resource. But then usage itself is difficult to quantify. There is a big difference between finding an article and using it in research and finding an article and doing nothing but print it out. References may be retrieved but the full text not pursued. The only way any statistics can be produced is by the database provider. They provide statistics but in may different formats. In consequence a body has been established called COUNTER – Counting Online Usage of NeTworked Electronic Resources (13). It says on their website: "Librarians want to understand better how the information they buy from a variety of sources is being used; publishers want to know how the information products they disseminate are being accessed. An essential requirement to meet these objectives is an agreed international set of standards and protocols governing the recording and exchange of online usage data. The COUNTER Codes of Practice provide these standards and protocols."

COUNTER currently provides two Codes of Practice, one for Journals and Databases and one for Books and Reference Works. In August 2008, Release 3 the valid Code of Practice for Journals and Databases went into effect (vendors have until July 31, 2009 to comply.) The current release for Books and Reference Works is Release 1.

COUNTER-compliant reports (often just called "COUNTER reports") are usage reports that are formatted exactly as defined in the COUNTER Code of Practice and use defined ways to count usage. When usage reports have the same kinds of data and are formatted the same way, they can be compared to each other and can be automatically retrieved into local systems The SUSHI Reports Registry lists the names used to make requests.

After a trial period, The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol (14) was officially published as a standard in 2007 (Z39.93-2007). SUSHI defines an automated request and response model for the harvesting of electronic resource usage data, using a Web services framework. It is intended to replace manual collection of usage data reports. In August 2008, a Standing Committee was approved to assume maintenance responsibilities of this standard, including encouraging the further use and adoption of this extensible, lightweight standard.

In the context of SUSHI, the COUNTER reports formatted in XML are the data which are requested

and delivered using the SUSHI protocol. Delivery of COUNTER reports via the SUSHI protocol is included as a requirement in Release 3 of the COUNTER Code of Practice. The implementation of the XML-based SUSHI protocol by vendors will allow the automated retrieval of the COUNTER usage reports into local systems, making this process much less time consuming for the librarian or library consortium administrator.

3.2 Identification

Retrieval of articles in journal issues is not so much of a problem. An article retrieved may display an item pulled off a database with a Digital Object Identifier (DOI). This is governed by practices which have already been set up but are the subject of the development of a standard: ISO/DIS 26324, DOI. (15).

If a journal is retrieved any display can lead through a hierarchy to find a particular article in a journal issue which may be stored at any URL. However there are holdings formats which can be used and these were developed many years ago by NISO and have been adopted by ISO. ANSI/NISO Z39.71 - Holdings Statements for Bibliographic Items (16) is the latest NISO standard superseding two earlier standards. ISO 10324:1997, Information and documentation — Holdings statements — Summary level (17) is based on one of the now withdrawn NISO standards.

These standards are intended to ensure that data on holdings is understandable to the human eye. There is also just published a standard for the implementation of holdings in XML: ISO 20775:2009 Information and documentation — Schema for holdings information (18). The schema is designed to cover the holdings of all types of resources, physical and electronic, all types of resource format such as printed text, visual images, sound recordings, videos, electronic media and resources published or issued once such as monographs or those published serially or in part.

The schema is primarily designed to be included in responses to queries. Two primary query types have been identified and targeted, based on availability and historical usage.

Although the schema may be used for reporting holdings to a federated metadata repository such as a centralized union catalogue, metasearch database such as Google or centralized document repository, this is not its primary focus. The focus of this schema is for interactive exchange of a combination of stable and dynamic information. Reporting and harvesting convey only stable information and other schemas are already in use for this purpose such as MODS (19), MARC21 Holdings (20) and the emerging ONIX SOH[3]. Most of these schemas include richer detail especially in relation to serial holdings. For this same reason, the schema is not intended to contain the detail necessary to predict new serial issues and claim missing serial issues.

How data is gathered and assembled to populate the holdings schema is also outside the scope of the standard. Data may be dispersed in several locations such as a union catalogue, local catalogue and a policy directory or repository. A variety of standards may be employed for this purpose including NCIP for local holdings (21), XACML (22) and LDAP (23) for policy, authentication and authorisation information and SRU (24) and Z39.50 (25) for all types of searching and retrieval.

The schema includes an optional section for identification and description of one or more bibliographic resources; however detailed resource

description is out of scope for this standard. The whole bibliographic resource section is optional so that the schema may be incorporated as a fragment within other XML bibliographic resource descriptions such as MODS.

An OpenURL as defined in Z39.88: the OpenURL framework for context dependent services (26) enables the transfer of metadata about an item (a journal article or book, for example) from a resource, where a citation is discovered (for example, an Abstracting & Indexing (A&I) database), to a link resolver. By providing a means to tell another system what something is, rather than where it is located on the Internet (the function of a normal URL), OpenURLs provide a means for link resolvers to take charge of directing users at particular institutions or organisations to appropriate, subscribed resources for the content, be they in electronic or print form.

This solves a critical problem for librarians: direct URL linking from one publisher's content to another's, including CrossRef DOI-based links, has the potential to lead users to resources that are inappropriate for them, i.e. to instances of content to which their institution does not subscribe. This results in users being refused free access to material because they have been directed by a provider to which the user's library has no subscription. In addition, where multiple subscriptions are held or a number of relevant access points exist, the librarian may desire to nominate one instance of the full text for the user rather than others (for example, should they be directed to the publisher's version or to one hosted by an aggregator?).

OpenURL linking not only improves the online working environment for library patrons by reducing the number of linking dead ends but it also – by improving content visibility – increases the usage of the library's licensed and subscribed materials and potentially reduces document delivery spend, all appealing outcomes for librarians.

The OpenURL linking syntax was first developed in 2000 at the University of Ghent, from which the first commercially available link resolver (Ex Libris' SFX) arose 2001. This linking syntax (known informally as Version 0.1) was, despite its unofficial status, quickly adopted by a significant number of content providers and library systems suppliers. At the same time, the syntax was earmarked for fast tracking to official approval by NISO. The NISO-approved syntax (informally known as Version 1.0, but officially as Z39.88), was released in 2004. It overcomes some of the limitations of the earlier syntax and is more extensible to other content types. It is therefore intended to replace the earlier syntax. However, the present reality is that both versions of the OpenURL syntax are in use in scholarly information today.

3.3 Licensing

Another area of identification relates to ensuring the user is in the set of people entitled to see a particular resource. There need to be rules on who qualifies in an institution, what about alumni for example.

NISO has developed SERU: A Shared Electronic Resource Understanding. The SERU Recommended Practice document (NISO-RP-7-2008) (27).

SERU offers publishers and librarians the opportunity to save both the time and the costs associated with a negotiated and signed license agreement by agreeing to operate within a framework of shared understanding and good faith. Among the issues covered in the SERU best-practice document are perpetual access, archiving, and interlibrary loan.

Publication of the SERU best-practice document in February 2008 followed a six-month trial use period, during which time librarians and publishers reported on their experiences using the draft document. NISO is in the process of producing additional materials to help publishers and libraries adopt a SERU approach, maintain a registry of participants, and continue to promote, educate, and plan for regular review and evaluation of SERU.

3.4 Preservation

A very important area for the digital library, particularly for those materials which are purchased but perhaps equally for material which is freely available but of equivalent value is the preservation of the material. Many materials are available under license. In the days of a printed periodical, when an issue was purchased you could keep it for ever. Initially when periodicals were digitised the licenses often were such that if you ceased to subscribe to a periodical you lost all the issues in electronic form that you had ever subscribed to. This has now changed and in the case of many journals for a small fee to cover administration of access you can keep the periodicals you have 'purchased' in the past. What happens if such material is withdrawn because a company providing access goes bankrupt or is taken over. A number of projects have been set up to deal with this. Many publishers allow users

to archive files providing only that the copies are not misused. In theory in many of these cases, these copies may be used if for example the system goes down. keeping these working can be a difficult task so the restructuring is not for the faint-hearted. However one organization LOCKKS has been set up. It stands for Lots of Copies Keeps Stuff Safe. The idea is that the more libraries keep a copy the safer it will be if the main source of the data breaks down in the future and the material is lost.. One of the main problems which beset preservation in ensuring that in the future systems will be able to read current formats (especially documents with combinations of text and image) so they are setting up procedures to avoid obsolescence, using standard formats such as GIF, Graphics Interchange Format, a standard developed by the World Wide Web Consortium,W3C (28).

4. Conclusion

Most standards that are developed begin as proprietary standards. Some are developed by foundations set up for one purpose such as DOI and COUNTER. Others like MARC came from the Library of Congress. NISO is active in standards development and ISO to a lesser extent though ISO often adopts NISO standards. It is easy to take for granted the influence of standards on the information world, but be assured that without the standards mentioned in this paper the retrieval of material from digital libraries would be much more complicated than it is.

References

1. **Sconul. Sconul** statistics on the web. London, 2008- http://www.sconul.ac.uk/
2. **Document management** — Portable document format — Part 1: PDF 1.7 PDF (ISO 32000-1:2008)
3. **Joint Steering Committee for Revision** of AACR .Anglo-American Cataloguing Rules 2nd ed., 2002 Revision: 2005 Update. Washington DC, ALA; Ottawa, CLA; London, CILI, 2005.
4. **Format for Information Interchange.** Geneva, ISO, 2008 (ISO 2709:2008)
5. **MARC21** is found at http://www.loc.gov/marc/

6. **Dierickx, H. and Hopkinson, A.** Reference Manual for machine-readable bibliographic descriptions. 2nd ed. Paris, UNESCO, 1981
7. **CCF/B** : the Common Communication Format for Bibliographic Information. Paris, UNESCO, 1992.
8. **Document management** — Portable document format — Part 1: PDF 1.7. Geneva, ISO, 2008 (ISO 32000-1:2008)
9. **Counter** –Counting Online Usage of NeTworked Electronic Resources http://www.projectcounter.org/
10. **Common Command Language for Online Interactive** Information Retrieval, Washington, DC, 2008 (Z39.58)
11. **Information and documentation** — Commands for interactive text searching. Geneva, ISO, 1993 (ISO 8777:1993)
12. **MARC21** Op. cit.
13. Format for information interchange. Op. cit
14. **Information interchange format.** Washington, DC, NISO, 2001. (Z39.2)
15. **Application Service Definition and Protocol** Specification. Washington DC, NISO, 1995 (Z39.50)
16. **Information and Documentation**. Information Retrieval (Z39.50). Geneva, ISO, 1995 (ISO 23950)
17. **COUNTER** Op. cit.
18. **Standardized Usage Statistics Harvesting Initiative** (SUSHI) Protocol. Washington DC, NISO, 2007. (Z39.93)
19. **Digital Object Identifier** (DOI). Geneva, ISO, to be published. (ISO/DIS 26324))
20. **Holdings statements for bibliographic items**. Washington DC, NISO, 2006. (ANSI Z39.71)
21. **Information and Documentation**: Holdings Statements: Summary level. Geneva, ISO, 2006 (ISO 10324:1997)
22. **Information and Documentation** — Schema for holdings information, ISO, 2009 (ISO 20775:2009)
23. **Metadata Object Descriptions Schema** (MODS). Version 3. Washington DC, Library of Congress, 2008
24. **MARC21 Format for Holdings Data.** 2000 ed., update 9. Washington DC, Library of Congress, 2008 http://www.loc.gov/marc/holdings/echdhome.html
25. **ONIX for Serials**: SOH: Serials Online Holdings -Version 1.0 (revised September 2005), Version 1.1 (June 2007). Washington DC, Editeur, 2007
26. **NISO Circulation Interchange** Part 1: Protocol (NCIP). Washington DC, NISO, 2008 (Z39.83-1). NISO Circulation Interchange Protocol (NCIP) Part 2: Implementation Profile 1. Washington DC, NISO, 2008 (Z39.83-2).
27. **OASIS** (Organization for the Advancement of Structured Information Standards) eXtensible Access Control Markup Language (XACML) Technical Committee. eXtensible Access Control Markup Language
28. **Internet Engineering Task Force.** Light Directory Access Protocol v3. Fremont, Calif., IETF, 2006. http://tools.ietf.org/html/rfc4510

About Author

Mr. Alan Hopkinson,Technical Manager (Library Services), Middlesex University, The Burroughs London NW4 4BT, UK