research information network

Creating catalogues: bibliographic records in a networked world

A Research Information Network report

June 2009

Table of contents

Acknowledgements

This report is based on work undertaken by Ken Chad Consulting and Sero and commissioned by the Research Information Network (RIN). The RIN and Ken Chad Consulting thank the many people who gave freely of their time and expertise during all stages of the report and are most grateful for their contributions.

Summary

Bibliographic records play a central role in enabling users to find, locate and gain access to books and journals. The records are created and enhanced at different stages in a supply chain from publishers, through a range of intermediaries, to libraries and then to end-users.

The digital revolution has brought changes in the processes through which records are created and made available for use along this chain. Each actor has its own motivations, aligned to its particular business model, in creating, adding to, using or re-using bibliographic data; and each uses models and formats that suit its purposes. These formats are then frequently modified to meet the needs of those further along the chain.

This report looks at how bibliographic records for content held by UK academic and research libraries are created and distributed, for printed and electronic books, and for scholarly journals and journal articles; and at how they are utilised by all involved in the supply chain, from the publisher to the final end user.

Bibliographic data plays a particularly important role for academic and research libraries. These libraries need good bibliographic data to fulfil their mission of supporting research, learning and teaching. They devote considerable resources to acquiring, managing and creating data, so that their users can find the content they hold, and so they can manage their stock and ensure it meets the needs of their users. But the established ways of achieving those ends are coming under increasing challenge from two related sources:

- a perception that these traditional processes involve unnecessary duplication of effort which could be reduced or eliminated, and
- a belief that new web-based, aggregated services, developed by a wide range of organisations, provide better ways of creating and sharing a more comprehensive set of high-quality records, as well as offering much more attractive services for end-users.

Academic and research library catalogues are not prominently visible in an online environment dominated by large-scale aggregations of information. Bibliographic data relating to significant amounts of the content they hold in physical form, and to the greater proportion of the material to which they provide online access under licence agreements, are not included in their catalogues. Users therefore make use of other services to discover and gain access to the information sources they need, even when those resources have been purchased and made available by the library. There is also increasing interest from Government in making the information generated in, and by, public sector organisations more widely available for re-use, to generate greater economic benefit, social gain, and improvements to public services.

Key findings

Duplication, gaps and missed opportunities

Our key finding is that the current arrangements for producing and distributing bibliographic data for both books and journals involve duplications of effort, gaps in the available data, and missed opportunities. In common with other research done, we find that there would be considerable benefits if libraries, and other organisations in the supply chain, were to operate more at the network level. However, there are barriers to making significant moves in this direction.

Recommendation 1:
Working together to find solutions
Our key recommendation is that all those involved in creating, distributing and using bibliographic data must work together to find creative, practical and sustainable ways to increase the efficiency of current systems and to exploit the opportunities for developing new services.

Printed books

It is for printed books that the supply chain for bibliographic data is most well-established, and for which the coverage in library online catalogues (OPACs) is most comprehensive – although not totally so. Data are created and shared along a chain starting from publishers, through to aggregators, library suppliers, bibliographic utilities, the UK legal deposit libraries, and individual research libraries. Each of these groups has its own set of motivations and each adds value in its own way. But there is much duplication of effort, and libraries are concerned that the records they receive are not always of the quality needed for themselves and their users.

Libraries are therefore spending significant resources in editing the records they receive, as well as adding data to meet their own local needs. Sustaining and developing individual catalogues for the more than 160 university libraries in the UK demands considerable resources. A shared catalogue for the whole UK higher education (HE) sector, with dynamic links to local holdings, could bring enormous benefits, in terms of reduced costs, of a more comprehensive coverage of both national and local holdings with better-quality records. It would also provide the potential for developing new user-focused services allowing them to remain relevant to their users and to compete with Amazon, Google and others.

Recommendation 2:
Removing the barriers to shared catalogues
Libraries should give serious consideration to the benefits to moving from standalone catalogues to a shared catalogue for the whole UK HE sector. A meeting should be convened of representatives of all the key stakeholders, including the commercial vendors, aggregators, JISC and other national services, as well as academic and research libraries, to explore how the barriers to a shared catalogue might be reduced.

E-books

E-books are an increasingly important part of the academic library landscape, but the arrangements for creating and distributing bibliographic records for them are not well-established. No organisation in the UK currently provides a comprehensive set of metadata for e-books. And as more players enter the e-book market, the task of identifying what is available, and what the respective access and pricing policies are, becomes increasingly difficult.

Metadata for e-books are created separately from those for printed books, and they do not pass through the same chain of aggregators and bibliographic utilities, or through the controls built into arrangements for the legal deposit of printed books. Both the content and the encoding of the data may not be of the same quality as for printed books. Even when librarians wish to import e-book metadata into their catalogues, they often find that the records are of poor quality. These difficulties have been exacerbated by confusions over the allocation of identifiers to the different versions of e-books; and we support the steps being taken by the International ISBN Agency and Nielsen, among others, to remedy this.

From the perspective of end-users, however, immediate access to the full text of e-books reduces the value of bibliographic records describing them, especially where access is not provided via the library catalogue (users may be directed to separate e-book interfaces). For all these reasons, libraries may decide that they cannot justify the effort required to ensure that their catalogues include comprehensive and high-quality records for e-books.

The problems arising in creation and distribution of high-quality records at this relatively early stage in the development of the e-book market need to be resolved as soon as possible. Otherwise the costs of the current inefficiencies will become increasingly burdensome as this market grows.

Recommendation 3:
Listings of high quality records for e-books
Publishers and aggregators should work together with other interested groups in the supply chain, and with librarians, to consider how to establish comprehensive listings of high-quality records for e-books, and to seek agreement on standards for the content and format of such records.

Recommendation 4:
ISBN for e-publications
Publishers and aggregators should support the work of the International ISBN Agency, Nielsen and others to ensure that each version of an electronic publication should have its own ISBN.

Scholarly journals

Scholarly journals are the single most important means by which scholars publish and disseminate the results of their work. Metadata for journals and their contents are critically important to publishers, librarians and researchers alike. But readers of journals are less interested in information about journal *titles* than in getting direct access to the text of the articles relevant to their work. Metadata in library catalogues, however, typically relates to titles and holdings, rather than specific articles.

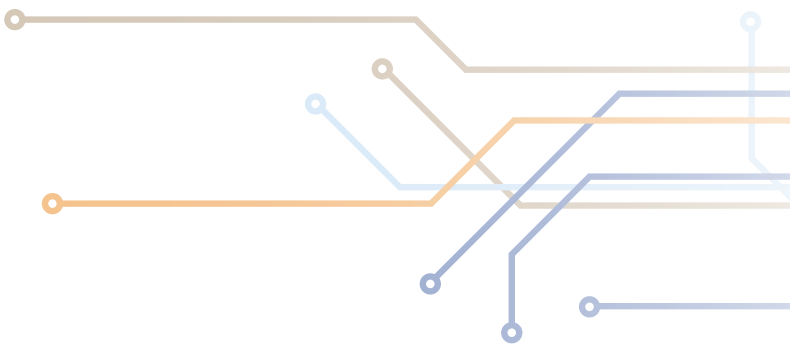
Even at the title level, the task of keeping data up-to-date is complex, and the quality of the records in library catalogues is patchy at best. To help solve some of these problems, the SUNCAT service was established in 2003 to aggregate title and holdings data from the catalogues of UK research libraries. Now that journals have moved to an almost wholly electronic environment, libraries are increasingly acquiring their title records from subscription agents and vendors who create lists from the data provided by publishers. Most libraries cannot justify putting effort into in-house journal cataloguing. Some do not load title data into their catalogues; they rely instead on linking (from lists on their websites, for example) to their vendors' hosted services.

Metadata for journal titles increasingly sits outside library catalogues, and the value of this metadata for end-users is coming into question. Lists of titles in a library catalogue or website are no longer a primary starting point for finding or following up a citation to a journal or article; and as linking technologies improve still further, the value of the title data in library OPACs will continue to decline. It seems likely, therefore, that the services being developed by commercial providers, alongside SUNCAT, will meet the needs of libraries for the foreseeable future.

Journal articles

Journal articles – now almost invariably in digital form - are overwhelmingly the most important category of information resources that researchers want to access. But the metadata relating to them are rarely found in library catalogues. Instead, users discover articles through a variety of services – abstract and indexing databases, publisher websites, Google Scholar and so on – and then gain access via a link resolver to an 'appropriate copy' of the full text, which will also be held outside the library.

Only rarely do metadata for articles flow into library catalogues, and libraries have not felt that local cataloguing effort would produce a useful service. So data and services flow from publishers, aggregators and other intermediaries direct to the user. Software has recently been developed to allow publishers to use RSS feeds to place Table of Contents (TOC) data into library catalogues. It is not yet clear how widely this kind of service will be taken up. But moves by publishers to make their metadata more widely available in a standard format could bring useful dividends to libraries and others in the supply chain, including end-users.



Journal articles in institutional and subject-based repositories

While metadata for articles are largely absent from library catalogues, they are absolutely essential for repositories. If subject-based and institutional repositories are to play a greater role in making articles more widely available, it is critical that users should be able to find materials stored in them, and also to ascertain the status of the copy or copies they hold.

Most of the metadata for the material deposited in institutional repositories is generated by the author, by a repository manager acting on the author's behalf, or added subsequently by a cataloguer. There are as yet few 'production' systems that draw in metadata from other sources. There is a clear need for better ways to get articles and their metadata into repositories, both to remove disincentives to researchers in depositing articles, and to increase the utility of repositories.

Recommendation 5:
Making metadata available
Publishers should make article-level metadata more widely available to third parties in a standard format, so that they can be harvested and utilised by aggregators, libraries, repositories and others.

The networked information environment

Libraries and their catalogues form a diminishing part of the global networked information environment. The growth of web-based services and the development of the web as platform mean that library and related services can and must be, and increasingly are, offered at a networked level, rather than by a single organisation.

Researchers and students are already using and relying on web-based services for search and navigation, as well as to download, create and modify bibliographic records and to share them with others. These and other services which make use of user-generated data in the form of ratings, tags and reviews, or recommender systems based on clickstreams, mean that the bibliographic records brought together in the catalogue of a single library are of decreasing value to end-users. These catalogues:

- usually provide reasonably high-quality and fairly comprehensive data about printed books, but often in ways that do not facilitate the aggregation and sharing of that data
- include only patchy data, of variable quality, about e-books
- provide data about journal titles that is again of variable quality, and also of declining utility for end-users
- rarely provide any information about scholarly journal articles, the single most important category of information resource for researchers, and
- seldom include records of the contents of the institution's repository.

While individual libraries still need good bibliographic records to

enable them to manage their holdings, the value of an individual library's catalogue for end-users is diminishing rapidly. If libraries, along with other key organisations in the supply chain, were to operate more at the network level, they would be better placed to:

- aggregate and make more productive use of data – including those supplied from organisations in the commercial and not-for-profit sectors – on a scale that more effectively meets users' needs, and further up the supply chain also facilitates the development of new services
- exploit their expertise to add value in meeting the needs of their users at both local and UK levels, and
- provide more comprehensive discovery services for all the kinds of content to which their users have access, whether it be in print, manuscript or digital form.

There are significant barriers to overcome in moving to the network level, even in relation to records for printed books. Mapping a way towards open platforms for sharing bibliographic data will require close attention to two related groups of issues:

- the need to develop a much clearer understanding of the motivations and the business models of all the players in the supply chain – publishers, aggregators, library suppliers, bibliographic utilities, the national libraries, libraries in the HE sector, as well as other players such as Google - and the incentives and constraints that are passed on through that chain, and
- the need for a much clearer definition of the standards and quality of the records required by users at each stage in the chain, of how those requirements can most effectively be met, and by whom. Without clear understanding and acknowledgement of the needs of all those who make use of the records at each stage, there is the risk that the current duplication of effort will continue, or even be exacerbated.

The RIN will work with the academic library community and other key stakeholders to raise awareness and understanding of the issues raised in this report, of the benefits to be gained by moving to new models, and of how we might overcome the barriers to achieving them.

Part 1: Context and background

Introduction

Bibliographic records play a central role in enabling users to find, locate and gain access to books and journals. They also enable all those in the books and journals supply chain to manage their resources effectively. The digital revolution has brought changes in the processes through which bibliographic records are created and made available for use along the supply chain, and utilised by end-users. It is not surprising that there is renewed interest in these issues; and we acknowledge the work being undertaken internationally, notably under the auspices of the Library of Congress and the Online Computer Library Center (OCLC), as well as recent initiatives by commercial organisations.

The focus in this report is on the bibliographic data created for and used by libraries in the HE and research sectors in the UK, and the communities they serve. Our findings may have wider relevance, but that is incidental to the work we have carried out and the conclusions we have drawn. We start by defining what we mean by bibliographic data and the purposes they serve, and then identify the main concerns expressed by the library community over current processes for creating and distributing such data.

Aims and scope

Bibliographic records are metadata: data about data. They provide information *about* "... printed and manuscript textual materials, computer files, maps, music, continuing resources, visual materials, and mixed materials. Bibliographic records commonly include titles, names, subjects, notes, publication data, and information about the physical description of an item" (Library of Congress, 2006). UK research libraries usually organise such metadata according to the Anglo-American Cataloguing Rules (AACR2) encoded for computer systems in the MARC (now usually MARC21) format.

Our focus is on bibliographic data (metadata) for printed and electronic books and for journals and journal articles. We have not considered the complex issues around authority records, nor do we address issues relating to the metadata for other kinds of materials such as maps or music scores, or for digital objects such as images, audio files or videos. These have been covered elsewhere, for example by the recent JISC report *Metadata for digital libraries: state of the art and future directions* (Gartner, 2008).

What are bibliographic records for?

Bibliographic data are created and enhanced at different stages in a supply chain starting with publishers, through a range of intermediaries, to libraries themselves, and then on to the end-users. Each of the actors in this chain has its own motivations, aligned to its particular 'business model' in creating, adding to, using or re-using bibliographic data; and each uses the format that suits its purposes.

Academic and research libraries seek to meet three broad purposes in creating and using bibliographic data:

- to enable their users to search for and discover the materials they hold
- to enable them to manage their stock of physical materials (printed books and journals for the purpose of this report). This requires localisation of records even if the record in other respects has originated from outside the library, and
- to enable them to ensure that their collections of both physical and electronic materials (taken here to encompass licensed as well as owned material) are appropriate for the learning, teaching and research needs of their institution.

These three purposes will remain as long as the HE system is organised around institutions that define their own mission and goals. But there is a perception that the system for producing and distributing bibliographic records, which has evolved piecemeal over time, needs attention. This perception is not limited to the UK, as witnessed by the Library of Congress Working Group on Bibliographic Control's report of November 2007. The report focused on the challenges and opportunities that the growth of web-based services provides, and envisaged that: "The future of bibliographic control will be collaborative, decentralized, international in scope, and Web-based. Its realization will occur in cooperation with the private sector, and with the active collaboration of library users. Data will be gathered from multiple sources; change will happen quickly; and bibliographic control will be dynamic, not static."

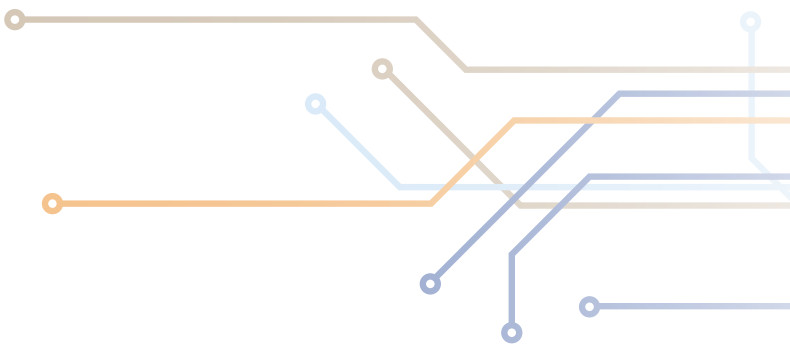
Defining challenges and working towards solutions

This report maps the processes through which bibliographic data are created and flow along the supply chain in the UK, to define the challenges now arising in relation to those processes, and to suggest how we might work towards some solutions. We do not seek to solve all the problems but rather to clarify the issues in order to set an agenda for further dialogue between the key stakeholders in the public, not-for-profit and commercial sectors. We believe that the challenges arise from two related sources:

- **Efficiency:** we examine the issue of the widespread perception of an unnecessary duplication of effort which could and should be eliminated, to free up resources for other work. We seek to distinguish between unnecessary duplication and local enhancement of records.
- **New web-based services:** these are having a profound effect on bibliographic services, arising from:
  - the ease with which bibliographic data can now be created, shared and used on a global scale
  - the increasing awareness in business and in government that 'opening up' all kinds of data for re-use can facilitate innovation, new services for users, and growth in the wider economy. *(Office of Public Sector Information, 2007)*
  - the popularity of commercially-based discovery and content services, particularly Google, whose business models allow them to deliver services free at the point of use
  - the growth in popularity of user-driven sites, often with a social networking dimension, including Open Library and LibraryThing, and
  - the growing interest from the library sector in the scope for harnessing data about user behaviour online to develop recommender systems similar to those provided by Amazon and others. See, for example, the JISC's TILE project at www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/tile.aspx

We believe that changes in how bibliographic records are created and shared could lead to new and improved services to users as well as cost savings. We seek to identify barriers to the development of such new services, and to open up the debate on whether and how these barriers can be removed.

Some but not all such new services will be user-led. Researchers and others are changing their methods of seeking and finding information with increasing speed. This brings the risk for libraries and other intermediaries that they will lose their relevance as information providers. The key challenge for them is to meet the changing needs of their users by applying their expertise and knowledge in new ways.



Part 2: Creation and flow patterns

In this section we survey the processes for creating and distributing bibliographic records in the UK, and identify the main pressure points in the current system. For convenience and analytic clarity we have distinguished between books and journals, between print and electronic versions and, with regard to journals, between journal titles and journal articles.

1. Printed books

1.1 Bibliographic record creation and flow patterns

Publishers

The creation of bibliographic data begins with publishers, who use them as the basis for sales catalogues and related information on their web sites. Producing metadata for library catalogues is not their mission and, with some small exceptions, publishers do not provide data direct to libraries. Rather, they channel the data through aggregators and other intermediaries who manage the necessary translation in format (e.g ONIX to MARC) and other issues in order to provide records geared to the library market

Intermediaries: aggregators

From the publisher, metadata move to 'aggregators' such as Bowker, Nielsen and Bibliographic Data Services (BDS). These companies add more data to the original records produced by publishers, including links to book-jacket images, blurbs, and information about possible interest groups or appropriate reading age. For aggregators, metadata are core business. They sell data and derive from them other services which they can also sell. Nielsen, for example, uses metadata to produce information on market trends. Its BookScan service produces, 'the official bestseller chart for the UK each week, collating real time sales from over 8,000 retailers representing over 90% of the UK book market'.

In the UK, Nielsen is the ISBN agency and it also provides a feed into the retail system used by major bookshops. Publishers therefore have a strong incentive to provide metadata to Nielsen, and while a basic listing in the Nielsen Bookdatabase is free, some publishers pay Nielsen to enhance their metadata.

Aggregators make the metadata supply chain more efficient by giving publishers a simple way for their metadata to be discovered and consumed by agencies further along the chain. They also seek to meet the needs of libraries by providing for them 'library quality' records. However, the evidence from our survey reveals that research libraries in the UK do not get their records directly from the aggregators. Instead these libraries take them from suppliers such as Dawson Books and Coutts or from bibliographic utilities such as Talis Base, the Research Libraries UK (RLUK) database, or OCLC; or they create them in-house.

Intermediaries: library suppliers

For library suppliers, such as Coutts and Dawsons, metadata are essentially a supporting tool for their core business of selling print and electronic content. They take feeds from aggregators and also direct from publishers. Some suppliers employ sizable teams of bibliographic staff to enable them to provide, along with the books they supply, 'library quality' records that can be loaded directly into the library's catalogue and Library Management System (LMS). In response to our survey, one supplier said that the take-up of such services in research libraries is still small.

Intermediaries: bibliographic utilities

Bibliographic utilities in the commercial sector, such as Talis Base, also take data from aggregators, while library-membership organisations such as OCLC and RLUK rely heavily on records contributed by their members and users. In both cases, bibliographic utilities charge their users for the data they provide, either as a direct fee, or as part of a fee for a wider group of services, or as a 'membership' fee.

The national and legal deposit libraries

The British Library creates and collects data from a variety of sources and makes authoritative records available through a variety of channels (some free, some charged) as part of its bibliographic services including the British National Bibliography (BNB) and other datasets. It outsources some of its record creation to intermediaries and currently sells records to individual organisations, aggregators, vendors of library management systems, library suppliers, and bibliographic utilities. Individual MARC records may be downloaded without charge from its online catalogue. In this way these records find their way to libraries either directly or through other channels.

The UK's five legal deposit libraries (including the British Library) have an agreement to ensure that each contributes full-level records for items received from a predetermined subset of the UK publishing output. They thus seek to ensure that the BNB, and services derived from it, include 'full-level' records for the whole of that output.

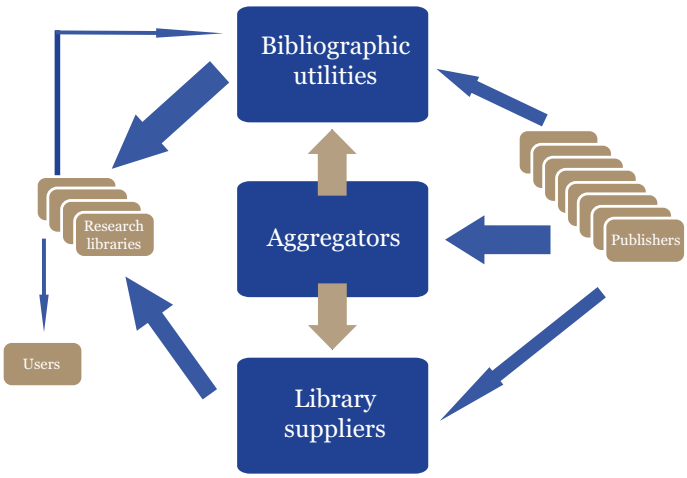
As noted above, the British Library outsources some record creation work. Pre-publication or Cataloguing-in-Publication (CIP) records for books are currently created by BDS. Book suppliers such as Coutts are also contracted to acquire non-UK books and to supply the associated records. The British Library also gets records under licence from sources throughout the international library community, including the Library of Congress and OCLC. Quality control standards are established in the contracts, and monitored by record sampling.

Libraries

Our survey of UK research libraries indicates that well over 80% of records for printed books are sourced externally rather than created from scratch in-house. Most research libraries - not just the legal deposit libraries - also *contribute* records for sharing via union catalogues or bibliographic utilities such as the RLUK database, Copac, Talis Base, and OCLC.

In addition to the relatively small percentage of records created from scratch, libraries edit and enhance the records they receive from external sources. Enhancements may include: subject classification; date; authority records to help ensure consistency (eg author name); and local data on holdings and location (which will usually sit outside the core bibliographic record).

Figure 1: Creation and flow pattern for books
(the larger the arrow the greater the flow)



1.2 From publishers to libraries: adding value

As bibliographic records move along the chain from publisher to library, value is added in a number of ways.

Aggregators
Aggregators add value by consolidating data, providing efficiencies of scale. They also convert data encoded for the 'trade' (usually now in ONIX format) into 'library' (MARC) format for consumption by bibliographic utilities, library suppliers and a few libraries. They may add data such as 'library' subject headings and authority data as well as enrichments such as reviews and images of book jackets. In essence, they have a simple business model: data, rather than content, are their business and they charge for them.

Library suppliers
Library suppliers' core business is to provide content to libraries, in print and increasingly in digital form. But they also add significant value to data by providing records that conform to the AACR cataloguing rules and by adding Library of Congress classification numbers. They help to ensure that the data can be integrated into the acquisition process through technologies such as Electronic Data Interchange (EDI), providing for example 'shelf ready' books that are serviced with spine labels and are accompanied by MARC records. Unlike aggregators, library suppliers use data to support their core business of providing print (and increasingly electronic) content; hence in some cases they are prepared to supply the data without charge.

Bibliographic utilities
Companies and organisations such as Talis Base, the British Library, RLUK and OCLC provide their subscribers or members with MARC records for import into their local LMS either online or in batch mode. They also provide a mechanism for sharing records, which reduces the amount of original cataloguing libraries need to do. They all charge for the data, though this may be covered in a number of ways. For OCLC the 'shared cataloguing' service it provides to its members is still the core part of its business. For Talis, it is essentially an 'add-on' to its core LMS business.

Libraries
In response to our survey, libraries explained that they add value by upgrading records to meet the standards they require, ensuring consistency in their records, and providing local data to improve discovery (through, for example, local classification numbers). Some libraries devote considerable efforts to editing Dewey or Library of Congress classification numbers in order to conform with local practice.

1.3 From publishers to libraries: achieving efficiencies and moving forward

The supply chain of bibliographic data for printed material has evolved over many years and the model itself is straightforward. But there are two areas of concern. The first is not so much complexity as apparent duplication of effort; the second is the perception by libraries that the content of the records they receive is not always of the quality they need.

In the UK HE sector, more than 160 catalogues require bibliographic data in one form or another. A major research library we surveyed employs nineteen staff, who edit each year around 16,000 records taken from external sources. Clearly the cost of enhancing records in this way is not insignificant, and we presume that there is believed to be a worthwhile return for the time and other resources spent on this work. In this study we do not explore those costs, but the recent RIN report, *Uncovering hidden resources: extending the coverage of online catalogues* (November 2007) highlights the backlog of work to be done before all the significant material held in UK libraries that could be of value to researchers can be readily accessed through online catalogues.

The balance to be struck between the effort expended in enhancing records for new acquisitions on the one hand, and in dealing with cataloguing backlogs or with retro-conversion of card catalogues into online form on the other, is clearly an important issue for some libraries. Editing records for new acquisitions is seen as an integral part of the continuous process of acquisition, but dealing with backlogs and retro-conversion tends to be seen in terms of projects that require additional funding.

Key issues for the academic library community are whether the current balance of effort is effective in meeting the needs of users, and whether the library catalogues of more than 160 universities, requiring locally-adapted bibliographic data, provide the best return on investment. The Library of Congress Working Group on Bibliographic Control in 2008 pointed to some of the problems associated with maintaining the status quo: “Redundant work means wasted resources. Time and money are spent redoing work that has already been done, rather than creating new records for materials not yet catalogued”. A catalogue shared across many institutions extending ideally to the whole UK HE sector or even beyond with dynamic links to local holdings, could bring a number of benefits:

- it would reduce duplication of effort and costs
- it would strengthen UK academic libraries in their discussions with the intermediaries about the quality of the records they receive from them
- it would promote ease of discovery through search engines such as Google. The Library of Congress Working Group notes, “Data that are stored in separate library databases often do not disclose themselves to Web applications, and thus do not appear in searches carried out through commonly used search engines. Such data are therefore invisible to information seekers using these Web applications, even though a library’s catalog may itself be openly available for use on the Web”, and
- it would facilitate the development of new services, such as the collective intelligence and recommender services pioneered by companies such as Amazon. These have made little headway in libraries largely because they depend for their effectiveness on large aggregated datasets Similarly, libraries have so far done little to exploit the potential for improving discovery services by making use of ‘contextual’ information relating to their users (course and year of study, for instance) or of information about users’ behaviour as revealed by their clickstreams (sometimes referred to as ‘intentional data’). Once again, the value of features that make use of such information increases as the dataset grows.

For all these reasons, shared or aggregated catalogues such as the RLUK database and Copac, or OCLC’s WorldCat, especially when linked to local holdings, can offer a user experience far superior to that of a single library’s OPAC, so can Google Books.

There are, of course, practical barriers to moves in this kind of direction. So long as acquisitions remain a local function, there remains a need for local bibliographic records of some kind; and where there are also related local shelfmarks, moves towards a common classmark, for example, could imply changing the physical arrangement of a library. The potential benefits far outweigh such difficulties, however, and we believe it is now urgent to consider how these barriers might be overcome.

Moving data to a web-based 'platform'

In the wider world we are seeing an increasing movement of data from local silos onto a web 'platform', where open 'application programme interfaces' (APIs) provide opportunities for innovation to deliver new services. Google maps, for example, is now commonly available and used in many non-Google services. Linked Open Data can be linked together to deliver new applications and insights.

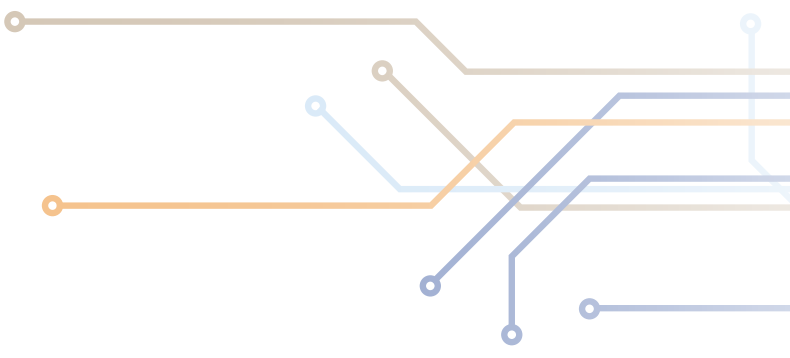
The move from traditional union catalogues to *platforms* remains at an early stage, and again there are practical obstacles to be overcome, including the need for integration with the functionality provided by library management systems, such as live local information about circulation and availability, and the ability to place requests. Initiatives like Jangle are working to enable interoperability of this kind. As products and services improve it seems likely that more libraries will adopt them.

'Early adopter' institutions have already exposed their bibliographic data to Google to make their collections more discoverable, with a simple a link to holdings. At present this is typically via OCLC WorldCat, which provides the switch through to the local catalogue to show availability. The motivation for these libraries appears to be not so much economies in cataloguing, but rather an attempt to retain users by delivering new services to meet the competition they face from Amazon, Wikipedia, Google and others.

1.4 Conclusion: sharing catalogues

The supply chain for bibliographic data that meets the needs of the HE and research communities in the UK involves publishers, aggregators, library suppliers, bibliographic utilities, the legal deposit libraries, and the broader range of research libraries. Each group of actors has its own set of motivations, and each adds value in its own way. But there is duplication of effort, and libraries are concerned that the records they receive are not always of the quality required to meet their needs and those of their users.

Libraries therefore spend significant resources in editing the records they receive, as well as adding data to meet their own local needs. Sustaining and developing over 160 plus academic library catalogues in the UK demands considerable resources. A shared catalogue for the whole UK HE sector could bring enormous benefits, in terms both of reduced costs and of the potential for developing new user-focused services to allow them to remain relevant and compete with Amazon, Google and others.



2. E-books

2.1 Bibliographic record creation and flow patterns

Google, Amazon and others have responded to and developed the appetite for e-books. From the user's perspective, they offer a number of attractions. Users of e-books can view the full text (or at least a 'snippet') through a variety of services, and more are on the way.

Virtually all printed books start in an electronic format, increasingly as 'full text' digital copies rather than simply as an electronic 'camera ready' version for print. Despite the obvious attractions to users, however, the e-books market has not yet developed significantly in the UK. Although one academic publisher, Taylor and Francis, told us that 80% of its publications are now available as e-books, we are still very far from the position in the UK where all books are published and available to readers in digital form. Both publishers and the HE community have an interest in investigating how the e-book market might be stimulated, especially for textbooks. This is a prime motivation behind the e-books observatory project commissioned by JISC in 2007, and a series of related studies commissioned in 2008 investigating business and licensing models.

For the purpose of this study, it is important to note that immediate access to the full text may reduce the value of bibliographic records for end users, especially since such immediate access is not yet always available through library catalogues (users may be directed to separate e-book interfaces). Metadata for e-books are, however, of importance to libraries in acquiring and managing their collections.

Metadata issues

The metadata issues raised by e-books are very different from those for print versions. A key immediate issue is that there is as yet no single comprehensive service enabling libraries – or anyone else – to discover whether a title is available as an e-book and, if so, where it can be obtained and on what terms it may be purchased or licensed. As aggregators add more e-books to their lists, however, they may be in a position to meet this need. But many issues remain to be resolved.

In 2003 the Gold Leaf report recommended that publishers should be urged to supply information about e-books to the providers of bibliographic databases; that intermediary services should work with the database services to provide comprehensive listings of e-books; that OPACs and other databases should be searchable by product and should provide links between printed and digital versions of the same book; that there should be better guidance on cataloguing of e-books and on ensuring the persistence of links by using Digital Object Identifiers (DOIs) or Uniform Resource Names (URNs); that metadata should include basic Dublin Core elements as well as publisher statement, bibliographic history and blurbs and abstracts to facilitate both selection and subject classification; that the use of ONIX to expose metadata for OAI harvesting should be explored; and that standards for a Rights Data Dictionary and Rights Expression Language should be developed. This remains a good summary of the issues still to be resolved today.

E-book identifiers

A problem identified in the Gold Leaf report is confusion about identifiers. An e-book may come in a variety of 'tradable products': in different formats, available from different platforms (the publisher's own website/platform, NetLibrary, Myilibrary, ebrary, etc), and available under different licence terms (for example, different restrictions on the amount that can be printed). Moreover, e-books are easy (technically at least) to disaggregate, into chapters or other subsets which could be licensed separately. This makes for a complex environment that taxes existing bibliographic standards: licence terms for example are not easily expressed or encoded in a MARC format.

In 2008 the International ISBN Agency restated its requirement that each different format of an electronic publication should have a separate ISBN. In August 2008 Nielsen stated that it would accept e-books for listing only where each 'tradable product' has a unique ISBN. Meanwhile, the issue of multiple identifiers is being addressed by Google through its Book Search API, and will be tackled also by the new Book Rights Registry to be established under the terms of the settlement between Google, the Association of American Publishers, and the Authors Guild.

Flow patterns

The creation and flow of bibliographic data and records relating to e-books originates, as for printed books, with publishers. But the flow to libraries takes a different route, since e-books do not have to be warehoused and physically shipped to libraries and retailers.

Publishers

E-books are not available through wholesale and retail outlets in the manner of printed works. Hence libraries acquire some e-books direct from the publisher. But publishers are not geared up to provide 'library quality' records, and so some use bibliographic utilities such as OCLC or Library of Congress to help to deliver records of the required quality. One publisher told us of the complexity and cost involved, "To meet the needs of library customers for e-books we have outsourced creation of MARC records - each customer wants their own MARC record standard. To get MARC records we first crawl the Library of Congress and if that fails it's outsourced to a specialist supplier who adds data such as subject categorisation [and] charges $15 per record". It is perhaps not surprising that the publisher also said, "MARC is too old and too inflexible and needs to be dumped. Libraries should accept XML/ONIX".

This indicates that metadata for e-books are not necessarily derived from the same processes and systems as those for printed books. The data appear to be recreated for the e-book, and they do not pass through an equivalent chain of aggregators and bibliographical utilities. Both the content and the encoding of the data may not be of the same quality as for printed books.

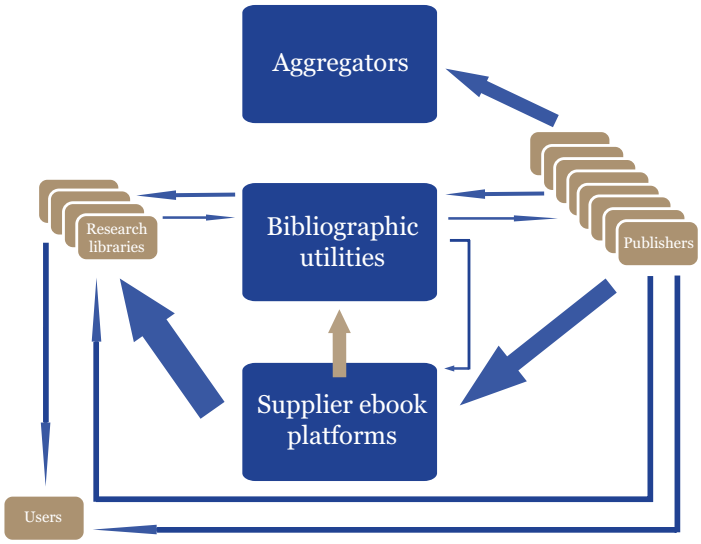
Library suppliers of e-books

Libraries purchase e-books (typically as a package of mixed titles) from e-book suppliers such as NetLibrary (a product of OCLC) or Myilibrary (a product of Ingrams), ebrary, Dawsons etc, or via JISC Collections. As part of the package, suppliers may offer the facility to import MARC records into library catalogues. The titles in the package to which library users have access may change over time, however, and libraries may not feel they can justify the resources necessary to keep the e-book records in their catalogue up to date.

Quality problems

The quality of the records provided for e-books has proved a significant problem for the JISC e-books observatory project. Hence a new National E-Books Observatory Catalogue Records (NEOCaR) project was established to provide libraries with a single download process for MARC 21 records for all the e-books licensed as part of the observatory. JISC Collections is now investigating a number of options to extend NEOCaR beyond the very small number of titles in the observatory project. One of the options is to ask publishers who have a JISC Collections agreement for e-books to place the MARC records for all their e-book titles into NEOCaR, thus creating a central searchable location for records. There is thus clearly a perception that the market is failing to deliver an appropriate service for academic libraries and that a new service such as NEOCaR is required to fill the gap.

Figure 2: Bibliographic data creation and flow pattern for e-books (the larger the arrow the greater the flow)



2.2 Adding value

Publishers

The slow development and market penetration of effective devices for reading e-books mean that there remains a large gulf between the content licensed to libraries for viewing page-by-page in a web browser on the one hand, and the sale of e-books for downloading in their entirety on a device on the other. For publishers, the value of providing metadata to an aggregator like Nielsen is thus not compelling at present, since it does not support a route to market. Moreover, since no legal requirement is yet in place to provide e-books to the legal deposit libraries, the bibliographic controls involved in the legal deposit process do not apply.

Intermediaries: aggregators, library suppliers and bibliographic utilities

No organisation in the UK yet provides a comprehensive aggregation of e-book metadata. Aggregators such as Nielsen cannot leverage their relationship with the retail market to motivate publishers to contribute data. But Nielsen's recent insistence, in line with the policy of the international ISBN, that each 'tradable product' must have a unique ISBN identity, may put them in a powerful position to fill the current gap in terms of bibliographic data services that are comprehensive in their coverage of what is available, from where and on what terms.

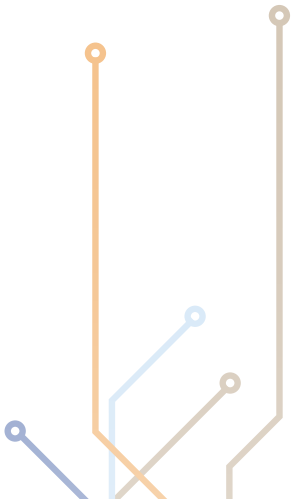
Coverage of e-books in bibliographic utilities is poor. OCLC has many e-books in WorldCat; but its ownership of NetLibrary may be a barrier to other e-book providers sharing data with OCLC. Both Copac and OCLC have announced that they are making use of a Google API to enable users to *link* from an OCLC or Copac record to the full text (or part of the text) made available through Google Books (Inside Google book search blog, 2009). The British Library licenses e-books for its research collections, but the absence as yet of a formal requirement for legal deposit of UK e-books means it has no opportunity to provide the kinds of bibliographical services that it does for UK printed books.

Google is clearly a major player in the e-book market because it enables users to find and gain access to significant quantities of content. It has digitized huge numbers of e-books – both in and out of copyright – and, following the recent settlement of its dispute with the Association of American Publishers and the Authors' Guild, it will be developing new services. No resolution of the issues around e-books is likely to be achievable without acknowledging Google's strong position in the market, along with the new Books Rights Registry.

2.3 Achieving efficiencies and moving forward

E-book publishing is at present much smaller in scale than print publishing; but it is growing. Traditional bibliographic utilities (Talis, Copac and OCLC), however, have not made significant inroads in e-book coverage. The danger is that bibliographic data creation and flow patterns for e-books will remain fragmented, and that the costs of this inefficiency will grow as the e-book market expands. But we are at a point where a solution to at least some of the problems outlined above might be found.

There is a need to engage all stakeholders who have an interest in e-books, including the commercial aggregators, in discussions to consider how we might most effectively move towards a single source of discovery. Any such move will have to be based on an agreement on the content and format of e-book records. It will also have to take full account of the business strategies and motivations of all the key players, and also of the likely development of the market in the light of Google's increasingly powerful position.



3. Scholarly journals: titles and holdings

3.1 Printed journals: creation and flow patterns

Scholarly journals are the single most important means by which scholars publish and disseminate the results of their work. Metadata for journals and their contents are critically important to publishers, librarians and users alike. But readers of journals are often less interested in information about journal *titles* than in their ability to secure direct access to the full text of the articles that are relevant to their work.

With a few minor exceptions metadata in library catalogues relates to journal titles and holdings (for example, Economic History Review: 1975-1987) rather than specific articles. But while the volume of metadata typically provided for journals is significantly less than for books, cataloguing journals can be complex. Libraries may get the initial bibliographic record from their bibliographic utility, but they have to spend considerable time creating and maintaining holdings data: keeping on top of new titles, cessations, changes and mergers is time-consuming.

Serials holdings and serial enumeration data are not reliably kept in a uniform format. MARC 21 was a step forward, but many libraries do not implement the holdings format, even when they adhere to MARC 21 in general. The British Library switched to MARC 21 in 2004, and LMS vendors began a similar switch for their customers. But by the end of 2008, Talis, for example, had not yet implemented the MARC 21 serial holdings format in its LMS system (Talis Alto) or its bibliographic utility (Talis Base). Indeed the transition to electronic journals and the potential of ONIX for serials may raise questions as to the value of further investment in enabling the MARC 21 holdings format.

These factors have created barriers to sharing data and enabling users to see which libraries hold which journals. To help solve some of these problems, SUNCAT was created in 2003-04 to aggregate serials title and holdings data from the catalogues of major UK research libraries. In addition to the data from

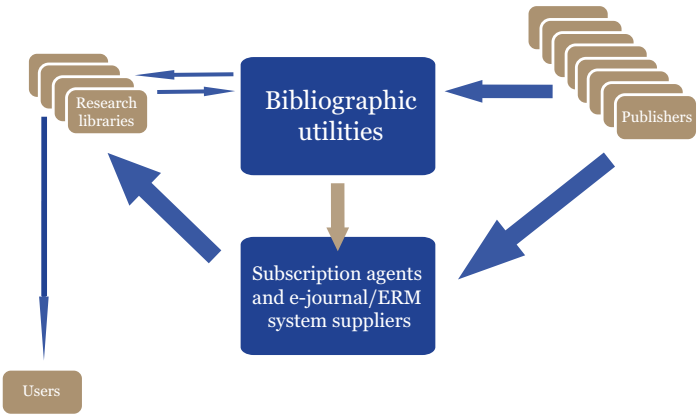
contributing libraries, SUNCAT includes records from the CONSER database, the ISSN Register, the Directory of Open Access Journals (DOAJ), and vendors such as Serial Solutions. CONSER records are generally high quality bibliographic records, and along with the ISSN records they can be downloaded in MARC format by libraries contributing to SUNCAT. Records from other contributing libraries will be made available for download in the future.

3.2 E-Journals: creation and flow patterns

Journals have moved much more quickly than monographs to an electronic environment, and hence traditional cataloguing practices are changing. Some records are still created locally and/ or downloaded from bibliographic utilities such as OCLC (where the CONSER database resides). However, libraries increasingly get their journal title records from E-Journal/ Electronic Resource Management (ERM) solutions vendors and subscription agents such as Serial Solutions, TD-Net, ExLibris and SWETS. These vendors typically provide batch files of MARC data to be loaded into local library catalogues.

Vendors and agents acquire title data from publishers, either through a feed of some kind, or by pulling down a list of titles from the publisher's website. The aggregated lists created in this way are of considerable value to libraries, which are increasingly outsourcing their journal cataloguing to services like these that enable them to provide (typically A-Z) lists of journals and links through to journal articles. With the move from print to electronic journals, it becomes hard for libraries to justify putting effort into journal cataloguing in-house. Indeed some academic libraries may not even load serial title data into their catalogues. They rely instead on linking (from the A-Z lists on their websites, for example) to the data on vendors' hosted services.

Figure 3: Creation and flow pattern for journal title records (the larger the arrow the greater the flow)



3.3 Adding value

In the process of aggregating lists of titles, vendors and agents add considerable value to the data they acquire from publishers. Thus in order to keep the various versions of titles and title abbreviations up-to-date, and variant ISSNs connected, vendors and agents exert "authority control" on journal titles, tying different versions the title together and to an authority file. This involves a considerable amount of work each month, but it makes link resolvers (see Section 4.2 below) work better because abbreviations in the source record can be translated into a different abbreviation in the target record.

3.4 Achieving efficiencies and moving forward

Although not perfect, the e-journal solutions provided by publishers, vendors, subscription agents and the bibliographic utilities have delivered for libraries and users both considerable economies and more effective services. Journal title metadata are shared in great part by the commercial vendors whose 'knowledge bases' are centralised repositories of metadata that are fed by publishers and agencies such as CONSER and Library of Congress.

While journal title metadata remains important to publishers, to intermediaries, and to libraries, the rapid shift from print to e-journals has brought with it declining interest in such metadata from the perspective of users. Lists of titles in a library catalogue or website are no longer a primary starting point for finding or following up a citation to a journal or journal article. A recent study reported that in 2005, the most likely starting point for users following up a citation were library web pages and OPACs, followed by specialist bibliographic and abstracting and indexing (A&I) databases; but by 2008, the A&I databases and the generalist search engines had gained in popularity to the detriment of all other possible starting points, and had eclipsed library web pages (Inger, S & Gardner, T, 2008). As linking technologies improve and become more widely adopted, the value to users of local cataloguing of journal titles in library OPACs will diminish further.

It seems likely, therefore, that the services being developed by commercial providers, alongside SUNCAT, will meet the needs of libraries for the foreseeable future, and that no further action is needed to stimulate the market.

4. Journal articles

4.1 Printed journal articles: creation and flow patterns

Offprints of printed journal articles are only exceptionally catalogued as independent entities in library collections. "Historically, access to the journal literature was a two-stage process. A user looked in one set of tools - abstracting and indexing services - to discover what was potentially of interest at the article level. Then they would have journal level access to the catalogue to check whether the library held the relevant issue." (Dempsey, L, 2006).

But print-only articles are an increasingly rare feature in the landscape: a recent survey (Cox, J & Cox L, 2008) found only ten small not-for-profit publishers who did not make their journals available online. Such journals are not now attractive to authors; for articles not exposed to the web are much less likely to be read and cited. We do not deal further with print-only articles in this report.

4.2 Electronic journal articles: creation and flow patterns

Metadata relating to articles in e-journals are typically stored outside library catalogues and are linked *to*, typically using link resolver systems provided by LMS or related suppliers. The development of link resolvers based on the Open URL standard has been crucial in facilitating access for users. Resolvers enable users who find journal articles through a variety of discovery services – such as A&I databases, publishers' websites, or more recently Google Scholar – to get access to the full text online in the form of the 'appropriate' copy for which their institution has purchased a licence. The ERM vendors and subscription agents who provide link resolvers search for metadata with federated searches, and feed them into their resolver. See CrossRef's *FastFacts* website at www.crossref.org/01company/16fastfacts.html for more information.

Therefore there is little flow of article metadata into library catalogues, although there are some exceptions to this rule. OCLC's WorldCat, for example, includes a significant set of article metadata derived from its own A&I databases, and via agreements made with a number of other publishers and aggregators; and data about journal articles may get into a library's LMS by being included on course reading lists. But more generally, users get to an article by discovering it in a database *external* to the library catalogue and then via a link to the full text, which is also stored outside the library.

To support linking, publishers support persistent Digital Object Identifiers (DOIs) through the cross-publisher organisation CrossRef which declares itself to be, "the citation linking backbone for all scholarly information in electronic form". Some publishers are also working to make their data more easily available. Oxford Journals, for example, have recently enabled feeds under the OAI-PMH protocol, explaining that such functionality offers third party aggregators and librarians a vastly improved way to get metadata records, "Metadata can be harvested at any time, as frequently as required". (Oxford Journals press release, 27 November 2006)

This kind of initiative is in line with the recommendations of a recent report commissioned by publishers, which encourages them to make their metadata more widely available:

"No-one can…predict where users will choose to start their research [and so] a publisher must actively back all of the navigational options for its readers and…collaborate with Google so that it optimally indexes the publisher's content; publish XML catalogues containing the metadata of its articles for library technology companies to harvest; support "deep-linking", OpenURL linking…promote its content to the key A&Is…and provide RSS feeds of recent content for other sites, such as portals, to pick up". (Inger, S & Gardner, T, 2008)

JISC has supported such developments with a range of projects under the PALS programme, notably the TOCRoSS project, which has developed software to allow publishers to use RSS feeds to place journal tables of content (TOC) data into library catalogues without human intervention. As yet, however, relatively few publishers have followed Oxford Journals' example by making their metadata freely available via RSS feeds or other means.

4.3 Adding value

With some exceptions, librarians have not felt that local cataloguing of journal articles adds sufficient value in return for the effort. Even uploading article metadata from external sources might cause significant additional burdens, since such metadata are typically based around a citation format rather than an AACR/MARC style. Making the two types of data congruent would demand significant resource without a clear benefit to users. Services such as TOCRoSS may change attitudes, but it is not yet clear how widely they will be taken up.

4.4 Journal articles in institutional repositories

A particular set of issues arises with regard to the metadata for articles deposited in institutional (and other) repositories. While metadata for articles are largely absent from library catalogues, they are absolutely essential for institutional repositories. Journal articles are a central part of the contents of repositories that aim to hold the research outputs of an institution or subject grouping, although there is considerable variation, and some confusion in terminology, surrounding the versions of the articles that repositories hold.

If institutional repositories are to play a greater role in making articles more widely available, it is critical that users should be able to find materials stored in them and also to ascertain the status of the copy or copies they hold: is it a pre-print (before or after peer review), or a version lacking formatting or copy-editing done by the publisher, or the published version, or some other version?

Currently most of the metadata in the majority of repositories is generated either by the author, by a repository manager acting on the author's behalf, or added subsequently by a cataloguer. Mediated services which deposit articles on behalf of authors may add metadata to facilitate searches across repositories. Federated searching is particularly important, since content (often with many authors from different institutions) may be deposited in several different repositories. Multiple deposit, of course, also brings with it duplication of effort in generating metadata.

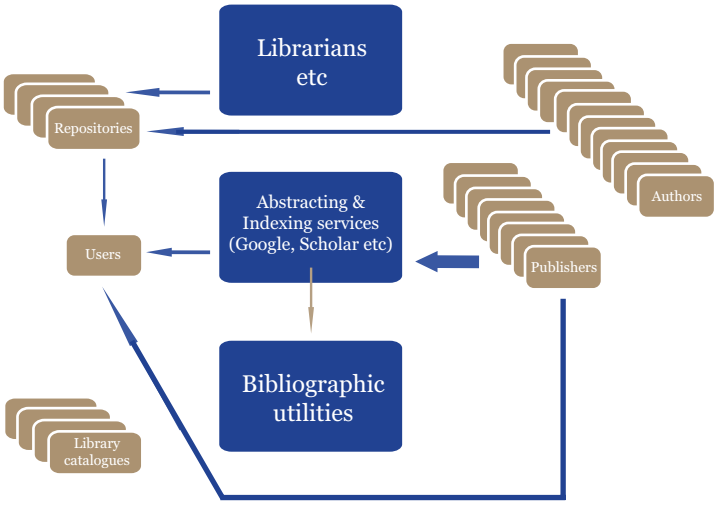
Author created metadata

Some argue that it is authors' responsibility to create the metadata at the point of deposit. But the effort required is one of the disincentives that prevent authors from depositing their articles at present. Authors are, of course, motivated to create some metadata when they submit an article to a journal, since publication is still the main means of securing academic recognition. That motivation does not yet apply to deposit in an institutional repository, although that may change if repositories succeed in becoming a significant route for access to and citation of articles and other research outputs.

Getting metadata from external sources

There are as yet few 'production' systems that draw in metadata from external datasets. But many universities are investigating ways to integrate their research publications database with the repository, or to use external sources of bibliographic information to pre-populate the repository. At the time of writing, JISC had recently invited proposals for a study of the actual and potential links between library OPACs and institutional repositories in the UK HE sector.

Figure 4: Creation and flow pattern for journal article records (the larger the arrow the greater the flow)



4.5 Achieving efficiencies and moving forward

There is a clear need for better ways to get article metadata into repositories. Some would argue that metadata creation is the *main* problem limiting the growth of repositories. At first glance it might therefore seem that improving the metadata workflow might be a route to solving the problem. JISC as well as commercial providers such as Thomson Scientific are working on ways to make online submission and deposit of articles fit more readily with researchers' workflows. But it seems likely that the bigger problem is ensuring that researchers can see real benefits as a result of depositing their work in institutional and other repositories.

Part 3: Wider issues

In previous sections we have focused on how bibliographic data are created and distributed, and how those processes might be improved. We have already touched on the scope for taking new approaches and exploiting the growth of web-based services that provide new opportunities but also new challenges, for research libraries in particular.

The web as a platform means that library and related services can and must be, and increasingly are, offered at a network level rather than by a single organisation – although the appropriate range of the network (regional, national, international, global) is itself a matter for debate. These developments provide scope for libraries and others to offer services and to add value for their users in new ways.

The networked information environment

Conventional libraries form a diminishing part in what is now a global networked information environment, based around the web. Global services like Google Books, Google Scholar and Open Library give users free access to a huge collection of resources, including bibliographic data. An environment characterized by information superabundance calls into question the traditional role libraries have played in shaping and ordering information and knowledge, based upon the bibliographic records they create and hold.

Libraries have typically ordered their records and their catalogues according to classification schemes such as those provided by Dewey or the Library of Congress (LC) and holdings in physical libraries are typically arranged around Dewey or LC classmarks. This ordering of information is common in other environments: retail consumer goods catalogues take essentially the same approach. In a digital environment, however, top down approaches where expert cataloguers decide what a book is about, and place it in a pre-ordained scheme of knowledge, may be less appropriate: users can reshape pre-ordained categories in ways that they find more useful for their own purposes. How libraries choose to organise their data and collections has no canonical authority: it is just one option among others.

In a sense, of course, this is not a new idea: researchers have always organised their collections of research materials in individual ways. But whereas once they used index cards, now they use personal bibliographic software or, increasingly, social networking sites. Through these, users can interact with and modify bibliographic records and share them with other users. It seems likely that the new Research Evaluation Framework (REF) will make it even more important both for individual researchers and for universities to ensure that they have high-quality bibliographic records of their publications, which they can customise for their own purposes. There is an urgent need also for discussions with publishers, ERM vendors and others to consider how comprehensive and accurate records can most effectively be generated and made available.

As for resource discovery, non-library organisations now offer a rich range of services by leveraging user-generated data in the form both of ratings, tags, and reviews, and of recommender systems based on clickstreams. In this world, library catalogues are no longer the sole or even the primary location for resource discovery; and the status of the bibliographic records created or held by libraries comes increasingly into question: where does the authoritative record reside, who creates it and who can modify it? In such a world, library catalogues may be regarded as no more than a record of locally-owned stock and licences; and users may question whether libraries can retain their traditional role as guardians of the quality of records and expert guides to appropriate research resources.

Re-use and innovation

The mood and rationale for change

The UK Government has expressed its desire to see public sector information (PSI) more widely available for re-use, on the grounds that such a move will bring economic benefit, social gain, and improvements to public services: "The availability of public sector information is essential to support the type of cumulative innovation required in a knowledge economy" (Pullinger, D and Sheridan, J, 2008). There are moves in the higher education sector to open up educational resources for re-use and adaptation.

In this context, the growing interest in making bibliographic data more widely and freely available coincides with related movements such as open access and more recently open data. In responding to the Library of Congress' Working Group on Bibliographic Control, Karen Coyle noted that "Open bibliographic data could bring significant benefits to the general public ..... as well as to other institutions and commercial developers." (Library of Congress, 2008)

The role of bibliographic utilities and libraries

The roles of libraries and of the bibliographic utilities – especially those in or close to the public sector – are thus coming under closer scrutiny. So are the nature of, and the constraints imposed by, their relationships with the various bodies, in the commercial sector and elsewhere, who provide them with data.

Moves towards opening up bibliographic data for free re-use are as yet in their infancy, but there have been some notable recent developments in addition to Oxford Journals' institution of data feeds already mentioned. LibLime, the US provider of open source software for libraries, launched in January 2009 a free browser-based cataloguing service, ǂbiblios.net, with a data store containing over thirty million records. The records are licensed under an Open Data Commons licence, and cataloguers can use and contribute to the database without restriction. Talis responded by announcing that they will provide ǂbiblios.net with, "data from the Talis Union Catalogue...including over 5 million bibliographic records catalogued by public and academic libraries in the UK." Some academic libraries are similarly making data available freely: the University of Huddersfield has released book circulation and recommendation data under an Open Data Commons licence. But the major bibliographic utilities have not as yet made similar moves.

Rights, licences and business models

Some have sought to argue that the 'facts' in bibliographic data cannot be protected by copyright or other means. But while a single fact cannot be protected, a collection of facts can enjoy copyright and/or database rights, or be protected under a licence agreement. So making use of records from an aggregator or bibliographic utility involves a real risk of infringement of copyright, of database rights or of the terms of a licence. The rights and business models of data producers can therefore act as barriers however to the free re-use of records.

At present the business models of the British Library, OCLC and RLUK do not permit the full and free sharing and re-use of

bibliographic data. This is a significant constraint on initiatives such as Open Library and the new ǂbiblios.net service described above. Some libraries have felt inhibited by their data licensing arrangement from contributing data to such services. As one senior librarian noted, "I don't want a situation where, every time a potential new export stream appears, we have to go round negotiating with all our data suppliers to ensure that they're happy with us feeding our records to this new target. It's simply too time-consuming and wasteful for every OCLC or RLUK member to have to deal with these issues". Existing licensing agreements with aggregators like BDS, Nielsen and others do not allow for such sharing and it is not clear what the financial incentives would be (on either side) for renegotiating these agreements.

RLUK and Copac

Copac is a JISC-funded resource discovery service based on the RLUK database, which covers a significant slice of the holdings of the major academic and research libraries in the UK. It is planning a number of significant developments to improve the service to users.

RLUK members can download to local library systems records held in the RLUK database; others must pay a fee to do so. Metadata contributions to the database are thus made available to the Copac resource discovery service, but they are not distributed freely to the wider HE community (or beyond) for download into library catalogues. This is a pity, since the records in the database are a valuable resource, whose value is being added to all the time (particularly at present through the Challenge Fund initiative which is adding a range of remarkable research collections from outside the RLUK membership). RLUK is currently considering, with JISC and others, the future development of the RLUK database and of Copac.

OCLC and WorldCat

OCLC provides an important *library centric* bibliographic database with a global 'web-scale' presence, illustrated by the fact that it is the default link ('borrow this book/ find this book in a library') from Google to local holdings. Its business model is based on cooperation, so members get records in exchange for their membership or contribution; and OCLC and is constantly seeking to develop new services. However, only a few UK libraries are OCLC members, and its business model sets constraints around the sharing and re-use of data. A significant recent development has been the establishment of a Review Board to consider its policies on the use and transfer of records from WorldCat. This follows an attempt to clarify and update those policies, which aroused a fierce round of accusations that OCLC was seeking to stop innovative use and reuse of library metadata, and thus promoting the marginalisation of library resources. It remains to be seen what new policies will emerge from the Review Board's deliberations, and whether the free sharing of data can be reconciled with the development of the OCLC business model.

The British Library

The British Library has been prominent in recent debates about copyright. Lynne Brindley, the Chief Executive, has argued that the current balance between private rights and the public domain is not working; that the public interest needs to be more actively protected; and that there is a need both for real innovation in business models and for legislation that is fit-for-purpose in the digital age (British Library press release November 2007).

These considerations are relevant to the British Library's role as a provider of bibliographic records. The British Library makes data available through bibliographic utilities such as OCLC, Talis Base, and Copac, and also through aggregators such as Nielsen, Bowker and BDS. Through its OPAC, records are also available to anyone without charge, but only on a record-by-record basis. Charging for bibliographic data is not central to the British Library's mission; but it generates significant income, and opening up the British Library's datasets would give rise to difficulties as to licensing agreements with suppliers as well as loss of revenue. In deciding

whether and if so how to change is current arrangements, therefore, the British Library cannot act alone: it will have to take full account of the implications for organisations in other parts of the supply chain, both those that supply it with bibliographic data, and those who depend on what it provides to them. Moreover, because it is impossible to predict what might be done with the data, and therefore what the benefits of re-use might be, it is hard to make a traditional business case. Nevertheless, the climate is changing, with initiatives such as ǂbiblios.net setting an example and putting increasing pressure on the British Library – as others in the supply chain - to review its stance.

The nature and scope of bibliographic records

One of the key issues for libraries and bibliographic utilities is the lack of agreement on the nature and scope of the data required to meet the needs of the different players in the supply chain. The standards for bibliographic records developed internationally by libraries and related organisations - by the International Federation of Library Associations (IFLA) through its FRBR framework, by the Library of Congress through its BIBCO and related programmes, and by the Dublin Core Metadata Initiative - all define extensive sets of fields. In the UK, the British Library (Bibliographic Standards) has described a vision of records with three layers:

- the core comprising the standardised description and the central authority data
- the special bibliographic data which will only be created when required, and
- the data needed for stock management.

None of the agents in the supply chain is meeting all these requirements at present. The lack of simple definition of what constitutes a 'core' record that might be freely shared and re-used is another significant barrier to establishing a more efficient system to meet the needs of libraries and their users.

Moving to a networked world: overcoming the barriers

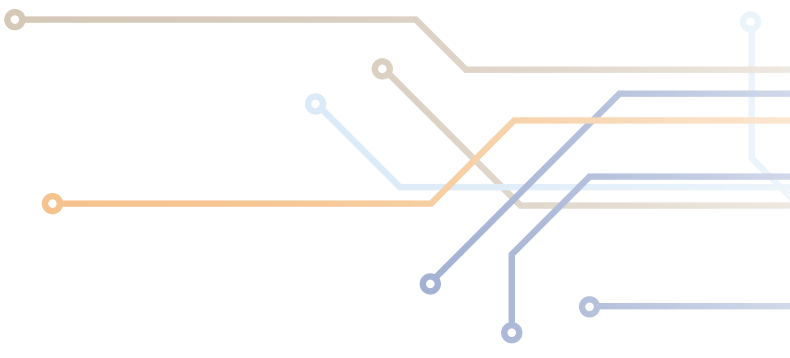
Moves towards a networked world in which bibliographic data are freely shared have considerable attractions. The pressures in favour of such moves are strong, both from the global web services, and from developments in public policy in the UK and elsewhere. It is in this context that OCLC has established the *Making metadata creation processes more effective* programme which aims to provide a common understanding of which data elements are critical to lead users to the resources held by libraries, archives, and museums, to information professionals responsible for the management of those resources, and to machine applications.

But the barriers to be overcome are also strong. The organisations that are *motivated* to provide free bibliographic data are ones for whom the provision of such data is not core business, including publishers and intermediaries such as library suppliers. For other organisations, open release of data could jeopardise not only current business models but also the overall supply chain in which value is added at every stage. The roles of all the organisations in the chain, especially bibliographic utilities and libraries would change significantly. And there are reservations in the UK library community about placing too much reliance on any bibliographic utility, still more a single global platform such as OCLC. Such reservations could be overcome, but establishing the necessary degrees of trust between libraries and utilities is not going to be straightforward.

Mapping a way towards open platforms for the sharing of bibliographic data will require close attention to two related groups of issues. First, we need to develop a much clearer understanding of the motivations and the business models of all the players in the supply chain, and the incentives and constraints that are passed on through that chain. This should encourage a discussion amongst the key groups of players –

publishers, aggregators, library suppliers, bibliographic utilities, the national libraries, libraries in the HE sector, as well as others, like Google – as to the models that might underlie platforms with interfaces to data open to others in both the commercial and non-commercial sectors. These interfaces would be a powerful incentive to innovation and to the development of new services for the benefit of all players, not least for end-users.

Second, we need a much clearer definition of the standards and quality of the records required by users at each stage in the chain, of how those requirements can most effectively be met, and by whom. Without agreement on these needs at each stage, there is the risk that the current duplication of effort will continue, or even be exacerbated. But UK academic libraries could – at the level of the first layer of bibliographic records defined by the British Library – agree on what is required for the core comprising the standardised description and the central authority data. They could then undertake quality control of such records on a shared basis, at a level of aggregation that could be national or global, leaving individual libraries to focus on adding data to records to meet strictly local needs.



Part 4: Conclusions and recommendations

In commissioning this work, the RIN was responding to a widespread perception that the current processes for creating bibliographic records and making them available to others are imperfect and inefficient.

Our key finding is that the current arrangements for producing and distributing bibliographic data for both books and journals do indeed involve duplications of effort, gaps in the available data, and missed opportunities.

Researchers and students are already using and relying on web-based services for search and navigation, as well as to download, create and modify bibliographic records and to share them with others. These and other services which make use of user-generated data in the form of ratings, tags and reviews, or recommender systems based on clickstreams, mean that the bibliographic records brought together in the catalogue of a single library are of decreasing value to end-users. These catalogues:

- usually provide reasonably high-quality and fairly comprehensive data about printed books, but often in ways that do not facilitate the aggregation and sharing of that data
- include only patchy data, of variable quality, about e-books
- provide data about journal titles that is again of variable quality, and also of declining utility for end-users
- rarely provide any information about scholarly journal articles, the single most important category of information resource for researchers, and
- seldom include records of the contents of the institution's repository.

While individual libraries still need good bibliographic records – which may come in a variety of forms – to enable them to manage their holdings, the value and utility of an individual library's catalogue for end-users is diminishing rapidly. If libraries, along with other key organisations in the supply chain, were to operate more at the network level, they would be better placed to:

- aggregate and make more productive use of data - including those supplied from organisations further up the supply chain - on a scale that more effectively meets the needs of users, and also facilitates the development of new services
- exploit their expertise to add value in meeting the needs of their users at both local and UK levels, and
- provide more comprehensive discovery services for all the kinds of content to which their users have access, whether it be in print, manuscript or digital form.

There are significant barriers to overcome in moving to the network level, even in relation to the bibliographic records for printed books. Mapping a way towards open platforms for the sharing of bibliographic data will require close attention to two related groups of issues:

- the need to develop a much clearer understanding of the motivations and the business models of all the players in the supply chain – publishers, aggregators, library suppliers, bibliographic utilities, the national libraries, libraries in the HE sector, as well as other players such as Google - and the incentives and constraints that are passed on through that chain, and
- the need for a much clearer definition of the standards and quality of the records required by users at each stage in the chain, of how those requirements can most effectively be met, and by whom. Without clear understanding and acknowledgement of the needs of all those who make use of the records at each stage, there is the risk that the current duplication of effort will continue, or even be exacerbated.

Recommendations

1. Working together to find solutions

All those involved in creating, distributing and using bibliographic data must work together to find creative, practical and sustainable ways to increase the efficiency of current systems, and to exploit the opportunities for developing new services.

2. Removing the barriers

Libraries should give serious consideration to the benefits to moving from standalone catalogues to a shared catalogue for the whole UK HE sector. A meeting should be convened of representatives of all the key stakeholders, including the commercial vendors, aggregators, JISC and other national services, as well as academic and research libraries, to explore how the barriers to a shared catalogue might be reduced.

3. Listings of high quality records for e-books

Publishers and aggregators should work together with other interested groups in the supply chain, and with librarians, to consider how to establish comprehensive listings of high-quality records for e-books, and to seek agreement on standards for the content and format of such records.

4. ISBNs for e-publications

Publishers and aggregators should support the work of the International ISBN Agency, Nielsen and others to ensure that each version of an electronic publication should have its own ISBN.

5. Making metadata available

Publishers should make article-level metadata more widely available to third parties in a standard format, so that they can be harvested and utilised by aggregators, libraries, repositories and others.

The RIN will work with the academic library community and other key stakeholders to raise awareness and understanding of:

- the need to develop a much clearer understanding of the motivations and the business models of all the players in the supply chain – publishers, aggregators, library suppliers, bibliographic utilities, the national libraries, libraries in the HE sector, as well as other players such as Google - and the incentives and constraints that are passed on through that chain,
- the need for a much clearer definition of the standards and quality of the records required by users at each stage in the chain, of how those requirements can most effectively be met, and by whom. Without clear understanding and acknowledgement of the needs of all those who make use of the records at each stage, there is the risk that the current duplication of effort will continue, or even be exacerbated, and
- the benefits to be gained by moving to new models, and how we might overcome the barriers to achieving them.

Glossary

Anglo-American Cataloguing Rules (AACR)
Designed for use in the construction of catalogues and other lists in general libraries of all sizes. The rules cover the description of, and the provision of access points for, all library materials commonly collected at the present time.

application programme interfaces (APIs)
Sets of routines, protocols and tools for building software applications.

British National Bibliography (BNB)
Records the publishing activity of the United Kingdom and the Republic of Ireland and as such is a measure of their intellectual output. This has traditionally included printed publications and more recently has been extended to electronic publications following the extension of legal deposit to this class of material in 2003.

CONSER database
Resides within the OCLC Online Union Catalog. CONSER members input, authenticate, and modify serial cataloging records on OCLC or contribute original records via FTP. Authentication is the process of approving the bibliographic elements in the record and providing for the record's availability through distribution services and bibliographic products.

Digital Object Identifiers (DOI)
A system for identifying content objects in the digital environment. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI name will not change.

Electronic Data Interchange (EDI)
Refers to the structured transmission of data between organisations by electronic means. It is used to transfer electronic documents from one computer system to another.

Electronic Resource Management (ERM)
Refers to practices and software systems used by libraries to keep track of important information about electronic information resources, especially internet-based resources such as electronic journals, databases, and electronic books.

Full-level records
Encoded in the MARC record as 'ukscp', which is a '042' field that the Program for Cooperative Cataloging (PCC) participants (BIBCO and CONSER) use to indicate that a record has been reviewed and authenticated. Code ukblcatcopy signifies that the British Library has used another organization's catalogue record essentially "as is" for its cataloguing, and that all name headings have been checked against the relevant authority file.

International Standard Book Number (ISBN)
A unique, numeric commercial book identifier based upon the 9-digit Standard Book Numbering (SBN) code.

JISC Publisher and Library/ Learning Solutions (PALS)
The aim of the PALS Metadata and Interoperability initiative is to facilitate collaboration between the HE/FE and publishing communities and develop practical solutions for metadata and interoperability.

JISC TOCRoSS: Table of Contents by Really Simple Syndication (TOCRoss)
The aim of TOCRoSS was to see if RSS could be used to automate the population of OPACs with details of journal articles, without the need for manual cataloguing, classification or data entry

JISC TILE project
The TILE project will contribute to the implementation of the JISC Information Environment and the JISC Libraries of the Future initiative by investigating developments in Library 2.0 services, within the context of developing a library domain model for the international e-Framework

Just Another Next Generation Library Environment (Jangle)
An open source project designed to facilitate API access to library systems

Linked Open Data
A relatively new concept describing the use of the web to connect data that were not previously linked, or to lower the barriers to linking by connecting data currently connected by other methods

MARC (now usually MARC21)
The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form, and related documentation. They provide the protocol by which computers exchange, use, and interpret bibliographic information. The data elements make up the foundation of most library catalogs used today

National E-books Observatory Catalogue Records (NEOCaR)
A JISC project that provides HE librarians with a single download process for the MARC 21 records for all the e-books licensed as part of the JISC national e-book observatory project.

Online Computer Centre Library (OCLC)
Not for profit computer service and research organization whose systems help libraries locate, acquire, catalog, and lend library materials

Open Archives Initiative (OAI)
Develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content.

ONIX
Both a data dictionary of the elements which go to make up a product record and a standard means by which product data can be transmitted electronically by publishers to data aggregators, wholesalers, booksellers and anyone else involved in the sale of their publications. ONIX was devised to simplify the provision of product information to online retailers by standardising the means by which information about the product was delivered and processed.

Online Public Access Catalogue (OPAC)
An online database of materials held by a library or group of libraries. Users typically search a library catalog to locate books, videos, and audio recordings owned or licensed by a library

Open data (OD)
An emerging term in the process of defining how scientific data may be published and re-used without price or permission barriers. Scientists generally see published data as belonging to the scientific community, but many publishers claim copyright over data and will not allow its re-use without permission.

Program for Cooperative Cataloging (PCC)
An international cooperative program coordinated jointly by the Library of Congress & PCC participants around the world

Research Evaluation Framework (REF)
HEFCE's funding and research assessment framework

Really Simple Syndication (RSS)
A family of web feed formats used to publish frequently updated works - such as blog entries, news headlines, audio, and video - in a standardised format

SUNCAT
The Serials Union Catalogue for the UK research community is a free tool to help researchers and librarians locate serials held in the UK

Uniform Resource Name (URN)
A Uniform Resource Identifier (URI) that uses the URN scheme, and does not imply availability of the identified resource. Both URNs (names) and URLs (locators) are URIs, and a particular URI may be a name and a locator at the same time

References

British Library (2009)
Lynne Brindley, CEO of the British Library, criticises debate on Intellectual Property as 'too focused on teenagers, music and consumer industries'- Balance in IP "not working". 30 November
www.bl.uk/news/2007/pressrelease20071130a.html
accessed 04.03.09

Cox, J & Cox, L (2008)
Scholarly publishing practice 3. ALPSP.
http://www.alpsp.org/ngen_public/article.asp?id=200&did=47&aid=24781&st=&oaid=-1
accessed 04.03.09

CrossRef (2009)
FastFacts: CrossRef's OpenURL Resolver
www.crossref.org/03libraries/16openurl.html
accessed 04.03.09

Dempsey, L (2006)
'The library catalogue in the new discovery environment: some thoughts' in *Ariadne*. Issue 48:30;7.
www.ariadne.ac.uk/issue48/dempsey
accessed 04.03.09

Gartner, R (2008)
Metadata for digital libraries: state of the art and future directions. JISC Technology & Standards Watch
www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf
accessed 04.03.09

Inside Google Book Search blog (2009)
Preview books anywhere with the new Google Book Search API. 13 March. http://booksearch.blogspot.com/2008/03/preview-books-anywhere-with-new-google.html
accessed 18.03.09

Inger, S & Gardner, T (2008)
How readers navigate to scholarly content. Comparing the changing user behaviour between 2005 and 2008 and its impact on publisher web site design and function
www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf
accessed 04.03.09

Library of Congress (2006)
Introduction to the MARC bibliographic record
www.loc.gov/marc/bibliographic/bdintro.html
accessed 04.03.09

Library of Congress Working Group on Bibliographic Control (2007)
Report on the future of bibliographic control www.loc.gov/bibliographic-future/news/lcwg-report-draft-11-30-07-final.pdf
accessed 04.03.09

Library of Congress Working Group on Bibliographic Control (2008)
On the record: Report of The Library of Congress Working Group on the Future of Bibliographic Control www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf
accessed 04.03.09

Office of Public Sector Information (2007)
The Power of Information: An independent review by Ed Mayo and Tom Steinberg. Command Paper Cm7157, 25 June.
www.opsi.gov.uk/advice/poi/power-of-information-review.pdf
accessed 04.03.09

Oxford Journals (2006)
Oxford Journals offers faster and better access to metadata records with OAI-PMH functionality. Press release, 27 November. www.oxfordjournals.org/news/2006/11/27/oxford_journals_offers_faster_an.html
accessed 04.03.09

Pullinger, D & Sheridan, J (2008)
Using semantic web mark up conventions to facilitate effective re-use of public sector information. Online Information Conference Proceedings

Research Information Network (2007)
Uncovering hidden resources: extending the coverage of online catalogues www.rin.ac.uk/catalogue-coverage accessed 04.03.09

Serial Solutions (2009)
360 MARC updates
www.serialssolutions.com/ss_360_marc_updates.html
accessed 18.03.09

Useful links

All accessed 04.03.09

BIBCO
www.loc.gov/catdir/pcc/bibco

Biblios.net
https://biblios.net

British Library's mission
www.bl.uk/aboutus/stratpolprog/redeflib/mission

British National Bibliography (BNB)
www.bl.uk/bibliographic/datalicensing.html

CONSER database
www.loc.gov/acq/conser/aboutcn1.html

Directory of open access journals
www.doaj.org

Dublin Core Metadata Initiative
http://dublincore.org/index.shtml

Functional Requirements for Bibliographic Records, International Federation of Library Associations (IFLA)
http://archive.ifla.org/VII/s13/frbr/frbr_current7.htm

ISSN Register
www.issn.org

Jangle
http://jangle.org

JISC ITT: the links between library OPACs and repositories in higher education institutions
www.jisc.ac.uk/fundingopportunities/funding_calls/2009/02/opacs.aspx

JISC national e-books observatory project
www.jiscebooksproject.org

LibLime
http://liblime.com

Library of Congress classification
www.loc.gov/catdir/cpso/lcc.html

Legal Deposit Libraries Act 2003
www.opsi.gov.uk/acts/acts2003/ukpga_20030028_en_1

NEOCaR
http://edina.ac.uk/neocar

Nielsen Book e-book listings policy statement
http://www.nielsenbookdata.co.uk/uploads/press/3NielsenBook_EBookPolicyDocument_Aug08.pdf

Nielsen Bookscan web site
www.nielsenbookscan.co.uk/controller.php?page=107

Office of Public Sector Information (OPSI)
www.opsi.gov.uk

OpenLibrary
http://openlibrary.org

PALS programme (JISC)
http://www.jisc.ac.uk/whatwedo/programmes/pals2/synthesis/pals.aspx

Promoting the uptake of e-books in higher and further education. JISC.
www.jisc.ac.uk/uploaded_documents/PromotingeBooksReportB.pdf

RLUK database members
www.rluk.ac.uk/node/307

SUNCAT
www.suncat.ac.uk

TOCRoSS project
www.jisc.ac.uk/whatwedo/programmes/pals2/tocross.aspx

UK ISBN Agency
www.isbn.nielsenbook.co.uk/controller.php?page=121

About the Research Information Network

Who we are

The Research Information Network has been established by the higher education funding councils, the research councils, and the national libraries in the UK. We investigate how efficient and effective the information services provided for the UK research community are, how they are changing, and how they might be improved for the future. We help to ensure that researchers in the UK benefit from world-leading information services, so that they can sustain their position as among the most successful and productive researchers in the world.

What we work on

We provide policy, guidance and support, focusing on the current environment in information research and looking at future trends. Our work focuses on five key themes: **search and discovery, access and use of information services, scholarly communications, digital content and e-research, collaborative collection management and storage.**

How we communicate

As an independent voice, we can create debates that lead to real change. We use our reports and other publications, events and workshops, blogs, networks and the media to communicate our ideas. All our **publications** are available on our website at **www.rin.ac.uk**

This report is available at **www.rin.ac.uk/creating-catalogues**, along with a supplementary notes document. Hard copies can be ordered via email **contact@rin.ac.uk**

Get in touch with us

The Research Information Network
96 Euston Road
London
NW1 2DB
UK

Telephone **+44 (0)20 7412 7946**
Fax **+44 (0)20 7412 7339**
Email **contact@rin.ac.uk**