

# Union Catalogues for the UK

## An RIN Expert Seminar



30 January 2007

### Introduction

This seminar was convened by the Research Information Network (RIN) with the object of considering the functions and purposes of union catalogues, and a) the desirability and b) the feasibility of moving towards the establishment of enhanced union cataloguing of library collections across the UK.

Recent RIN survey work has demonstrated that researchers across all disciplines continue to make significant use of library catalogues alongside other discovery services, and that they value both comprehensive and specialist catalogue services. So how can development of the catalogues and related services we have at present improve what is on offer to researchers, and what are the roles of the various stakeholders involved in providing or facilitating these services?

The seminar brought together 30 experts from a range of bodies – in the commercial as well as the public sectors – concerned with developing and delivering cataloguing and related services for libraries and end-users. The aim was to try to develop a better understanding of what we have now and how it might be improved; how end-users might benefit from such improvements; and how the improvements might be achieved, by whom, and at what cost.

Twelve short papers were circulated in advance of the seminar, and discussion was grouped around six core sessions:

1. Mapping the key players and who is doing what in the UK; and an overseas perspective from the Netherlands
2. Issues in establishing architectures for aggregating bibliographic records
3. Future developments for COPAC and SUNCAT
4. The role of union catalogues in delivering a web-scale library presence
5. Issues of coverage and scope
6. Technological and other issues from a JISC perspective

Both the short papers and the PowerPoint presentations made by some of the participants are available on the RIN website at [rin.ac.uk/union-catalogues](http://rin.ac.uk/union-catalogues)

This report highlights the key issues discussed at the seminar, and which need to be addressed in finding the best ways to develop services further so that they more effectively meet the needs of different groups of end-users.

## Presentations and Discussion

### *1. Mapping the Key Players, and an Overseas Perspective*

Michael Jubb from the RIN introduced a paper that sought to identify the key organisations and groups currently involved in producing bibliographic and related records about the holdings of UK research libraries, and in providing discovery services that enable researchers and others to find out about such holdings, both at title and at article levels.

The key message was the map is very complex; and that the flows of information between the key players, and the relationships between the various services they provide, are even more complex. So there are questions as to whether the current complexity could be reduced, and whether services could be improved through more co-ordination. More immediately, would both librarians and end-users be helped by a clearer map of the current players and their activities, roles and responsibilities?

Inge Angevaare from the **Koninklijke Bibliotheek** (KB) in the Netherlands spoke about how the Dutch union catalogue developed from the joint cataloguing system established by PICA in the 1970s; the relationships between OCLC-PICA and the libraries in the Netherlands; the current situation with the union catalogue; and prospects for the future as the system sought to adapt to technological change and also to changes in user expectations.

Inge stressed that aggregating libraries' metadata in the union catalogue has brought important benefits to Dutch users; but there are important issues to consider for the future. She suggested that in an information world dominated by Google and other major players, an important role for libraries is to focus on meeting the needs of clearly-identified groups of users - whether they be defined by geography or by subject interests – for aggregations of metadata that are relevant to their interests.

National approaches remain important, not least because nations constitute communities of interests, and national governments provide funds and have a policy-making role; but for many groups, aggregations from across the world, provided in a web-based environment, could be more appropriate.

Issues that arose in discussion included

- The need to focus on the needs and expectations of end-users, particularly with regard to the interfaces between them and the information they seek
- The need also to take into account the needs of librarians in producing records and gathering them from other sources; and the potential for freeing up some of their time and other resources for other purposes if current systems could be made more efficient and effective, avoiding duplication of effort
- The need to address the issues posed by the large quantities of material in libraries that remain uncatalogued
- The potential for increasing the usefulness of catalogues by providing not only bibliographic records, but also information such as book jackets, reviews and so on; and the possibility of reaching national deals with the providers of such information, so that it can be included in union catalogues
- The need for co-ordination to avoid duplication of effort, but also the risks involved in large-scale planning that might undermine more local initiatives in what will remain a complex environment, with many different players and communities of interest.

## *2. Issues in establishing architectures for aggregating bibliographic records*

Gordon Dunsire, from Strathclyde University, spoke to a paper setting out the various aggregation methods, physical and virtual; and the importance in all cases of deduplication, consistent and coherent display of results, and interoperability. Variations in metadata structure and content have a significant impact on interoperability; and repeated aggregations of metadata in both physical and distributed union catalogues add to the complexity, since each aggregation may itself change the metadata format or content. Different aggregations have different communities, purpose or focus, and mapping source data to aggregations is highly complex.

Paul Watry, from Liverpool University, drew on the outputs of the **Cheshire digital library project** and the distributed architecture of the **Archives Hub**. He stressed that any aggregation has to face the issue of the varying quality and format of bibliographic records and related information, and diverse metadata vocabularies. Natural language processing techniques should be used to generate enhanced records, and to help with the challenge of deduplication. And text and data mining techniques should be used to help guide users – librarians as well as researchers - to the information they want.

Issues that arose in discussion included:

- The scope and granularity of the information to be included in union catalogues (tables of contents as well as traditional bibliographic records; reviews and commentaries?)
- The merits of **Functional Requirements for Bibliographic Records** (FRBR) as distinct from **Dublin Core** approaches, not least in tackling the issue of duplication of effort. It was suggested that the aim should be to ensure that the record for a *work* in the FRBR model should be created only once
- The possible role of unique identifiers
- How to ensure that any union cataloguing approach is scaleable, given the complexity of the underlying systems and of the relationships between the key players

## *3. Future developments for COPAC and SUNCAT*

Sean Dunne from **MIMAS** spoke to a paper prepared by Julia Chruszc (who unfortunately could not be present at the seminar) about **COPAC** and current development plans. These include additions to content from up to 17 new library collections; providing additional data in the form of review, TOCs, biographies etc; linking to the **English Short Title Catalogue** (ESTC); and increased functionality brought about by the move to an XML version of the database. There are lots of opportunities to develop the service further; and a key issue is ensuring that resources are available to allow the service to take up those development opportunities.

Fred Guy from **EDINA** spoke about **SUNCAT** and key issues that were arising as it grew and migrated from a development project into a national service. There are essentially two sets of services and interfaces: first, search and browse facilities for researchers, which could be developed further as part of the **JISC information environment**; and secondly facilities for librarians to download records. Data quality remains a significant issue in handling the records created by a large number of libraries; and there is obvious potential for links to access services and to other discovery services such as TOCs.

Issues that arose in discussion included:

- The need to build on what has been achieved by COPAC and SUNCAT; but to consider also whether and how they might be integrated as the basis for a national union catalogue
- The need also to consider integration with TOC services (whether **ZETOC** or some development of it, or some other service(s))
- Whether there is a case for ensuring cover of information that is not formally published but nevertheless deposited in repositories
- How to ensure the most effective exposure to Google and other search engines
- The need to distinguish between roles, responsibilities and activities relating to data creation, to providing views of that data, and to developing and providing services utilising that data for various purposes and audiences

#### *4. The role of union catalogues in delivering a web-scale library presence*

Paul Miller from **Talis** set out the ideas around the Talis platform, and how it might be used to insert libraries into the flow of information and activities employed by different groups of end-users. Sharing, aggregating and exposing data – along with the capability to manipulate and build applications around the data – could open up enormous opportunities. We need to think not so much in terms of a single union catalogue, but of as large as possible an aggregation of data through sharing. Various views, applications and services could be built upon such an aggregation by a variety of bodies to meet different needs.

Robin Murray from OCLC-PICA stressed the need to aggregate on a large scale to achieve an effective presence for libraries on the web (aggregation of supply), and pull from the web for library services. Optimising library presence on Google search lists depends on very large aggregations, which themselves depend on the use of FRBR. Union catalogues are therefore essential not only as aggregations of content, but as a mechanism for aggregating demand for library services, and as a switch which then redirects users to the most appropriate library service to meet their requirements.

Issues that arose in discussion included:

- The need to encourage ‘promiscuous’ sharing of data, and to overcome the barriers which are preventing it at present
- The need to think carefully about the services that libraries – and academic libraries in particular – can most effectively offer in an increasingly web-based world, and the implications of Web 2.0 developments.

## *5. Issues of coverage and scope*

Derek Law, from Strathclyde University, argued that we should build on what we have already in COPAC and SUNCAT, extending them to cover more collections, addressing the de-duplication issue, and adding TOCs and article-level searching. But adding *content* to the existing services, through retroconversion and tackling the large quantities of material held in libraries that are not yet catalogued is, he argued, the key issue to address.

Caroline Brazier, from the **British Library**, agreed that we need to find smarter ways of dealing with the historical legacy on uncatalogued materials; and to widen the categories of material that are included in catalogues (to include sound, audio and other digital objects, for example). There should be a focus on unique content, but we should also consider the role that collection-level descriptions can play.

Mike Mertens, from **CURL**, suggested that we should focus not just on the needs of researchers, but also of other groups. If the concept of the bibliographic 'long tail' is to be effectively realised for the UK knowledge economy, a virtuous circle of improving the visibility, access to, usage and integrity of all kinds of specialist collections would be essential. This process would be of benefit to all researchers, not just HE-based ones, as well as librarians, digitisers, and conservationists.

Issues that arose in discussion included:

- Agreement on the importance of adding content to existing catalogues through retroconversion and finding smarter ways to deal with the backlogs of uncatalogued material. A current **RIN study on retroconversion** should help to establish priorities in this area.
- Varying views on the usefulness of collection-level descriptions and how they are best created
- Researchers' desire not so much for information about an item, but for the item itself; and discussion of whether the focus should be more on extending the range of digital content, and search and retrieval functionality related to that content.
- The need to understand more about user behaviour before making decisions about investment in enhanced catalogue services; the usefulness of commercial services such as Amazon for tracking usage of material, user behaviour and also for tracing out of print or rare material.

## *6. Technological and other issues from a JISC perspective*

Rachel Bruce from **JISC** followed earlier speakers in urging the need to build on COPAC and SUNCAT, and also to build relationships with other services such as **Intute**. There have been significant changes since the UKNUC report of 2001, in the growth of Google and other search engines, developments in service-oriented architectures, and the emergence of registries to drive

There is a need for an assessment of possible architectures for joining services together, moving away from **Z39.50** to **Search/Retrieve via URL** (SRU) or open search approaches. And that should be accompanied by work on a rigorous business case (can't WorldCat or Google do it all?) and of related organisational issues.

## Summary of Key Issues and Conclusions

There was broad agreement that the seminar had been useful in identifying a core set of interrelated issues that needed to be considered further in order to find the most effective way forward:

1. **Adding content to existing catalogues and services.** Huge quantities of material remain uncatalogued, or catalogued only in print or on cards; and there was general agreement that addressing the backlog is a key priority. How it is tackled is important: we need to build on what has been achieved through the **Research Support Libraries Programme (RSLP)** and **Full Disclosure** initiative, establish priorities and avoid duplication of effort.
2. **Building larger-scale aggregations of data.** Researchers and other users, including librarians, want to be able to use large-scale aggregations as well as more specialised services. Aggregations thus need to be as large as possible, but available for view at various levels, from the national to the more specialised and local.
3. **Sustaining and developing UK-focused services.** Data and services can and should be shared across national boundaries, but services that can be developed and sustained at UK level remain valuable and important.
4. **Sharing data and exposing it to others.** Effective aggregation depends on willingness to share data as widely as possible. We need to find ways of enabling libraries and other bodies involved in creating bibliographic and related data to realise its value through sharing and exposure to third parties.
5. **Focusing on the behaviour and the needs of end-users.** Different groups of users – librarians, researchers and others – make use of catalogues and other finding aids in different ways, and we need to learn more about this as services develop; and to try to ensure that what is offered to them meets their needs.
6. **Establishing a web-scale presence for libraries and their services.** Inserting libraries into the flow of information and activities employed by different groups of end-users, and achieving pull from the web for the services of individual libraries, is key to the future of libraries of all kinds.
7. **Providing more kinds of information.** Evidence suggests that users value the bringing together in one place of information in the form of reviews, commentaries, author biographies etc alongside bibliographic information. We need to think about Web2.0 approaches to enable users to generate added value here.
8. **Operating with different levels of granularity.** There is clear evidence that researchers and others want article-level and chapter-level information to go alongside serial and book-level records. There were some differences of view about the value of collection-level records.
9. **Moving from discovery to delivery.** Researchers and others want to go directly from discovery to access on the desktop, and we need to facilitate this wherever possible.
10. **Reducing the complexity of the current flows of information between the key players.** It was agreed that this complexity is a barrier to effective sharing of data. There was discussion about the scope for co-ordination, and how that might be achieved without stifling local initiatives. Some participants argued the case for reducing the current complexity by adopting more top-down approaches at national or international levels. But it was acknowledged that there were significant risks in such approaches.

11. **Avoiding duplication of effort.** Valuable time and resources are currently used in creating data and records that have already been created by others. It was agreed that the aim should be to create bibliographic records once only.
12. **Improving data quality.** The variations in metadata structure, format and content have a significant impact on data quality, duplication of records, and interoperability. Some participants argued strongly for the adoption of approaches based on FRBR rather than Dublin Core.
13. **Building on COPAC and SUNCAT.** It is clearly important to build on the investment in and the achievements of COPAC and SUNCAT. But we need to consider how to achieve greater integration between them, along with TOC services and Intute.
14. **Exploiting new technologies.** Text and data mining, as well as Web2.0 technologies, have the potential to transform the services that can be provided using bibliographic and related data; and we need to exploit that potential to the full.
15. **Developing an architecture of functions.** We need to be able to describe much more clearly than we can at present both the current position and where we should like to be. We thus need to develop an architecture of the roles, responsibilities and activities relating to data creation, to providing views of that data, and to developing and providing services utilising that data for various purposes and audiences; and of the relationships between them. JISC is interested in undertaking or commissioning work in this area, based on the Clax report on the JISC-funded discovery services, and in collaboration with other interested parties.
16. **Developing a vision.** As well as an architecture, there is a need for a clear vision, underpinned by a set of principles and realistic objectives agreed by all the key players
17. **Next Steps.** Participants agreed that the seminar had been useful in bringing together many of the key players to discuss how to develop better services for both librarians and end-users; and that it would be useful to hold a further meeting to consider how best to take forward thinking on the issues that had been discussed.