Clumps or . . . organised access to printed scholarly material: outcomes from the third MODELS workshop

LORCAN DEMPSEY and ROSEMARY RUSSELL

ABSTRACT

The MODELS (MOving to Distributed Environments for Library Services) project is based around a series of five workshops. The third of these, 'Organising access to printed scholarly material', proposed a co-ordinated approach to providing access to a managed, distributed bibliographic resource. This article has two main ambitions: firstly it explores the influential outcomes of the third workshop, and secondly, it places this discussion in the wider MODELS context.

1. Introduction

1.1 MODELS motivation

In our world there are information resources.[1] These might be books or images or bibliographic records or web pages or journal articles or some other type of resource. Typically such information resources exist in collections. These collections might be databases, web sites, document supply centres, multimedia documents or libraries. In one perspective, the Internet itself is a collection. Such collections are also, of course, information resources, and collections may contain other collections.

Information resources may be categorised in various ways: they may be digital or paper-based, they may be simple or compound, they may be static or dynamic, they exist at different levels of granularity and aggregation.

Some useful distinctions can be made.[2] Information resources may be data or metadata, they may be final objects of interest in themselves or they may be data which assists in the identification, discovery, selection, location or documentation of resources. Catalogues are collections of metadata which describe collections of books. Web crawlers provide collections of rather terse metadata which describe network information resources.

Metadata services were developed independently across different domains, where a domain may be loosely defined by some combination of discipline, format and professional sector. Abstracting and indexing services typically describe journal articles. But they support discovery and selection, not location. Typically, inclusion criteria are subject and quality based: such services have no one-to-one correspondence with any collection of information resources. A library catalogue supports discovery, selection and location.

Typically, there is a relationship between a catalogue and a collection, and the catalogue tries to mirror some of the relationships between resources within the collection. An Internet 'search engine' supports discovery and location. And so on. Collections are managed within different curatorial or professional traditions, which have evolved differing technical and descriptive frameworks. For archives, museums, libraries, geospatial data repositories, and many other areas there are a variety of approaches between which there are growing points of contact and comparison.[3]

In a wider context metadata – at a variety of levels – is seen as central to mature networked information services. Users, human or automated, need to be able to discover the existence of resources, establish whether they are fit for intended use, establish the terms and conditions attached to their use, be assured that they are what they purport to be, and so on. In time, metadata will be pervasively disseminated throughout the network to describe people, services, software components and data, in support of self-describing, dynamically reconfigurable distributed systems and services.

Information resources, or collections of resources, may be grouped into 'resource spaces' depending on the protocols used to access, retrieve or manipulate them. Information resources may be searched and results retrieved using Z39.50, SQL, and so on. Information resources may be requested using ARTTel, http, BODOS (BIDS), one of a number of other approaches, and may be returned in print or electronic form. The physical use of libraries involves rather a different set of protocols involving going to shelves, browsing and so on.

Collections will also have different terms and conditions associated with their use. They are embedded in particular organisational and business practices, which may impose additional technical requirements on any networked solution, charging and copyright management for example. (In its current phase, MODELS has not focused on business issues, which are central to mature information environments.)

Access to resources will be typically mediated by a layer of software, or 'middleware', which shields the user from having to know in advance the characteristics or existence of all potentially relevant resources and from the mechanics of the complex interactions that may provide a service in response to any particular requirement. Below, we look in a little more detail at the functions such middleware could provide in the context of distributed catalogue access. Typically a user would use a 'thin' client (maybe a web browser) to interact with a 'home server' (or 'broker') which presents a managed information landscape based on the user's entitlement.[4] The abstraction at which this 'landscape' is presented will depend on the richness of the underlying middleware.

The MODELS project involves the examination of metadata and the collections they describe. Metadata and collections currently reside in multiple disjointed print and digital resource spaces, and in different business environments. MODELS seeks to explore some of the issues involved in reducing this disjointedness and moving towards more unified access to information systems and resources.

240

1.2 Project background

MODELS (MOving to Distributed Environments for Library Services) is a UKOLN initiative supported by the Electronic Libraries Programme[5] (eLib) and the British Library Research and Innovation Centre (BLRIC).[6] It is motivated by the recognised need to develop an applications framework to manage the rapidly multiplying range of distributed heterogeneous information resources and services being offered to libraries and their users. Without an appropriate framework, use of networked information will not be as effective as it should be. MODELS is providing a forum within which the UK library and information communities can explore shared concerns, address design and implementation issues, initiate concerted actions, and work towards a shared view of preferred systems and architectural solutions.

Electronic Libraries Programme funding supports a technical consultancy with Fretwell Downing Informatics, which advises on protocol and related issues. ELib funding also supports four workshops; BLRIC funding supports the fifth. UKOLN activities are jointly funded by the Joint Information Systems Committee (JISC) and BLRIC.

1.3 Project lines

To allow progress to be made, the project has partitioned the problem into five project lines. An invitational workshop is being held for each line. These inevitably have overlapping concerns and none will deliver a universal view, but we hope that by progressively working towards a picture that incorporates the insights of each, we are helping to contribute to future developments. At the same time, each is of self-standing interest. Each workshop has its specific outcomes, while, in parallel, a consolidated general framework will be proposed and distributed in report form in 1997. This will not propose any definitive solution, but will outline various of the building blocks which will contribute to distributed library services and explicate some of the issues to be addressed as they are assembled into real services. The MODELS approach has been found very useful and may be extended to cover related issues after the current phase. MODELS is advised by a small Steering Group and each workshop has a small planning group with appropriate representation.

It will be seen that the lines have been chosen because they present issues to services and systems developers at the moment and because they provide useful ways of highlighting general issues of moving to distributed library systems. It is not suggested that there are not other application areas demanding attention.

1.3.1 Article discovery and request

This first line looked at the discovery, location and request of journal articles. It is especially interesting because of the highly fragmented and variable levels of metadata content (abstracting and indexing services, catalogues, . . .) and the hierarchical nature of the information resource (title, volume, contribution). Accordingly, many of the necessary issues were raised in this line. This

241

line was useful in that it oriented future discussions and set the scene for the consideration of multiple services. It directly led to proposals for further studies and meetings, to be funded by JISC, which, at the time of writing, are the subject of a call for proposals.[7] Recommendations from the workshop are available on the UKOLN web server.[8] This workshop was chaired by Richard Heseltine, Director of Academic Services, University of Hull.

1.3.2 *Metadata for network information objects*

The second line looked at current approaches to metadata for network information objects. This built on previous work with the Dublin Core – a common set of metadata elements, and introduced the Warwick Framework, a container architecture for aggregating metadata. The outcome of this work, jointly carried out with OCLC, was reported in the *D-Lib Magazine*.[9]

1.3.3 *National resource discovery: organising access to printed scholarly material*

This line examined the problems of providing access to existing catalogue data based on the heterogeneous, fragmented resource that currently exists in the UK. The focus was on discovery, not on request or delivery of materials. It introduced business and organisational considerations which are likely to influence developments. A significant outcome was the recommendation for a National Agency for resource discovery. The British Library and JISC agreed to fund a scoping study for such an agency and after a competitive tender, the study is being carried out by a consortium of CERLIM (the Centre for Research in Library and Information Management) at the University of Central Lancashire and Fretwell Downing, with support from Geoffrey Hare (Essex County Librarian and Chair of EARL Council) and Index Data (a Danish software company). This workshop was also chaired by Richard Heseltine.

1.3.4 *Integrating access to resources across multiple domains*

User interests do not naturally fall into compartments which neatly align with existing curatorial or professional sectors. There are user interests which may be satisfied by documentary resources, time-based media resources, geospatial resources, and so on. They may be interested in museum objects, in archival materials, in books, and so on. It quickly became apparent that museums, archives, galleries, electronic text archives, and others responsible for selecting and organising parts of the intellectual record have shared concerns and this workshop looked at some of these issues. Conclusions are being formulated in the light of follow-on activities at the time of writing. This workshop was chaired by Chris Rusbridge, Director of the Electronic Libraries Programme.

1.3.5 *Managing access to a distributed library resource*

Recent developments have encouraged renewed attention to resource sharing at various levels. The current library systems environment does not support unified access to a library resource distributed across several libraries with

heterogeneous systems. This project line will examine some of the infrastructure which would support effective resource sharing, with a special focus on public library developments. The workshop takes place in April 1997, and will again be chaired by Richard Heseltine.

2.   Workshop on organising access to the printed scholarly material

This workshop was held over two days in July 1996. It involved 42 invited participants, selected as stakeholders and experts in the field. MODELS workshops generally consist of a series of presentations and discussions, with breakout sessions on the first afternoon to identify prominent issues. The final plenary session on the second morning consolidates issues and suggests a way forward, with a set of key recommendations.

The scope of this third workshop was described as follows:

Library catalogues represent an important resource discovery tool for printed scholarly material. However, a user who wishes to discover the existence or location of a particular item may have to individually visit, in an unguided way, a variety of catalogues. There are also significant bibliographic resources which may not be available for him or her to use. Catalogues include individual library catalogues, union catalogues and significant national services.

The aim of this workshop is to move towards a shared view of the technical requirements for unified national resource discovery systems in this area, acknowledging the organisational and business environments in which services now operate or are being developed. It will identify technical and organisational impediments to national resource discovery services and discuss ways in which they might be overcome. The primary focus will be on the higher education environment, with due reference to other sectors.

The focus of the workshop is on discovery. Other issues, such as requesting, are being dealt with in other parts of the MODELS project.

3.   Discussion at the workshop

There was general agreement that the current bibliographic resource was fragmented and poorly articulated and that services built on some more organised combination of existing resources were desirable. At this stage of development, it was recognised that a national union catalogue was infeasible for technical, organisational and service reasons. A distributed approach which provided unified access to existing catalogues was desirable, although a *fully* distributed option providing parallel access to all catalogues was undesirable for reasons of scale.

To clarify discussion, the notion of 'clump' was introduced. A clump is an aggregation of catalogues. A clump may be 'physical' where it has a continuous physically aggregated existence. Example of physical clumps are COPAC

243

(the CURL OPAC), the BL OPAC (made up of BL catalogues), the SLS and BLCMP union catalogues, Viscount or Unity. A clump may also be 'virtual', where the records from the participating catalogues are not physically brought together. How closely coupled the members of a virtual clump might be was seen as a discretionary matter, depending on the particular service scenario involved. The relationship might be entirely dynamic or user-defined, or it might be determined by long-standing service agreements among a group of service providers.

In breakout sessions and subsequent discussion some desiderata for an organised approach were outlined. These cluster into User issues and Catalogue issues and are presented below.

User issues
Ease of access; Consistency; Reliability; Availability; Quality; Relevance; Free at the point of use (?); Authoritative; Choice; Completeness.

Catalogue issues
Predictable behaviour (consistency in indexing, etc.); Include collection-level descriptions; Capable of being aggregated in helpful ways; Terms of availability for discovered material; Sustainable; Service standards.

There was a recognition that a controlled environment was necessary if desiderata were to be addressed.

4.   Standards

What standards would be necessary to allow the construction of distributed national resource discovery systems? Areas where standards were identified are listed below.

Standards
User interface (Z39.50/Java/windows, WWW); Referral services (WHOIS++, LDAP, ?); Search and retrieve (Z39.50); Collection level descriptions (?); Service levels (?); Authentication (?)

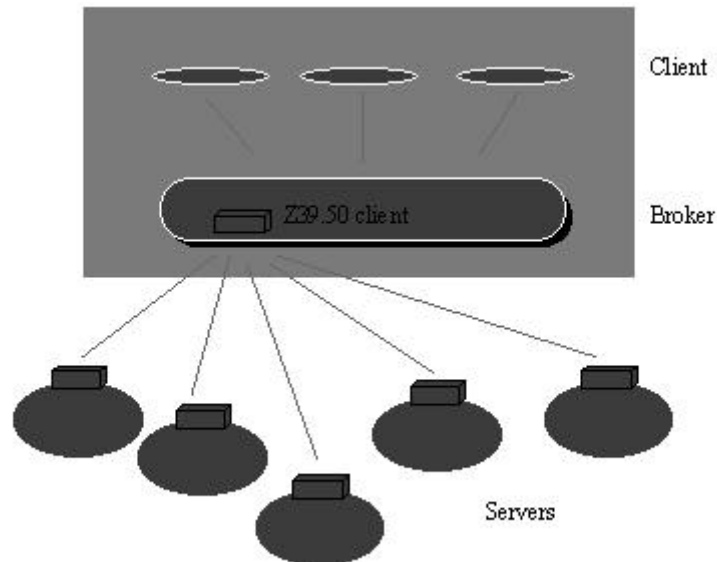These are considered in more detail in following sections.

244

Figure 1. The three-tier architecture of client, broker and servers

4.1   User interface and general architecture

A three-tier architecture was presented. A 'thin client' (typically a web browser or windows interface) was likely to be the preferred user instrument. A 'broker' service gives access to various services. The broker might be more or less rich. Minimally, for example, it might offer Z39.50 client functionality. It might open multiple Z39.50 sessions and consolidate the results. It might offer other services (request, etc). It might offer referral or directory services: pointing the user to potentially relevant resources to search. The future importance of object technologies, Java in particular, was noted as a way of delivering applications to the desktop.

It was anticipated that 'broker' services would be provided at various points in a service chain and would provide different levels of service. The broker might be a desktop application; a local campus service; a gateway to a commercial service or to a national service or set of such services; and so on.

4.2   Collection descriptions

There was an assumption that users would sometimes need to have collection-level descriptions, however they are provided. This was for a number of reasons. In general it aims to reduce search parallelism and redundancy. On some occasions a user would wish to do a higher level search to identify catalogues of potential interest. In some cases, a catalogue might not contain item level descriptions for some collections (e.g. a slide collection); a collection-level description might be better than no other representation. Collection level

description might form the basis for defining 'virtual clumps', or allow users to navigate sensibly through available resources; they could also help determine the likely value of a visit to a library. It should be noted that there is not necessarily a one to one relationship between collections and catalogues. All the items in a collection may not be represented in a catalogue. A catalogue may describe resources from more than one collection; more than one catalogue may describe the resources of one collection.

Several approaches to collection description were mentioned, including Conspectus and the Z39.50 profile for access to digital collection. A complementary approach was also briefly discussed, involving the generation of 'centroids', inverted index style representations of database content, which could be shared and searched in various ways.

4.3   Search and retrieve

There was general agreement that there was no alternative to Z39.50 here.[10] To meet the stated desiderata, however, implementors would have to additionally agree on particular options (e.g. mappings between attribute set elements and local indexes) to ensure predictability of service and consistency of results. Other implementation issues would also require attention (e.g. communicating availability data, non-delivery of MARC records to avoid record re-use, and so on). 'Virtual clumps' could be created by searching several resources and consolidating the results.

4.4   Referral services

Collection-level descriptions need to be made available for searching in some way. Again, a variety of approaches are possible. They could be served up through a directory service (WHOIS++, Lightweight Directory Access Protocol, . . .); they could be made available through the Z39.50 Explain service; they could be collected in a central database; and so on.

4.5   Service levels

It would be useful if libraries had consistent ways of describing service levels. As a minimum, whether and under what conditions items are available for loan need to be noted.

4.6   Authentication

This is an issue of considerable importance. Current network information environments are hampered by the requirement for users to individually log on to different services. The workshop noted some interim solutions and noted discussions about more fully distributed services.[11] There was agreement that a national framework should be supported, but an acknowledgement that this was a considerable organisational and development issue.

246

5.   Organisational issues

To be effective, it was important that developments proceeded within a well-understood and ordered framework which imposed no unnecessary overhead.

To participate in a 'national discovery system' it was decided that individual catalogues should meet some minimum criteria. These should be formally described in a Catalogue Profile, and providers would register the Profile with a National Agency. Initial components of such a profile, as developed at the workshop were:

- technical profile (relating to use of Z39.50);
- collection profile;
- service levels (under what conditions, if at all, would discovered items be made available; other services).

A National Agency would perform several functions:

- registration;
- profile management services;
- promotion and awareness;
- facilitate and liaise.

It was recognised that such an agency might have a broader remit in relation to other collections of scholarly materials (archives, electronic texts, museum objects, etc) and that it should be inclusive so long as this did not dissipate its focus. The Agency should have a truly national role and not be a creature of Higher Education alone.

6.   Current developments

6.1   Clumps

At the time of writing, a call for proposals[12] has just been released under JISC's new Committee for Electronic Information's (CEI) development programme, which includes provision for development work in this area. CEI plans to fund a small number of pilot virtual clumps. These could be regional or subject based. It is also hoped to include a pilot cross-domain searching demonstrator.

At the same time, the term 'clump' is entering more general currency and several working groups are looking at some of the practical issues of supporting clumps.

6.2   National Agency study

A scoping study for the National Agency is due to be completed by the end of April 1997. This has been jointly funded by JISC and the British Library. The study will collect opinion about whether such an agency would be useful, what roles it might carry out, and so on.

Program *vol. 31 no. 3*

6.3   The MODELS UK Interoperability Z39.50 profile

A 'profile' defines agreed options and use of the protocol. To meet some of the user and catalogue issues identified by the workshop it was thought useful to develop a profile. A group of Z39.50 developers was convened and a profile is in final draft form. This is available through the MODELS web pages.[13]

7.   Towards the future

This article has largely focused on the technical issues surrounding distributed access to catalogues. However, the interest shown in 'clumps' and surrounding issues meshes with a renewed interest in resource sharing and extended access. The recommendations of the Anderson Report, local government reorganisation and other factors point to a need to develop integrated technical frameworks to support higher level collaborative activity. It looks as if the library community in the UK is about to begin to experiment with genuinely distributed services.

References

1. There is an interesting terminological question here. 'Object' is useful, but is probably best not used to avoid confusion with particular meanings in an object-oriented context. A 'resource' is necessarily vague, but in a bibliographic context, we have the precedent of 'item' which has often been used without any precise referent.

2. Lorcan Dempsey and Maria Heijne. Scientific information supply – building networked information systems. *The Electronic Library*, vol. 14, no. 4, August 1996, pp. 317–332.

3. Lorcan Dempsey and Rachel Heery (with contributions from Martin Hamilton, Debra Hiom, Jon Knight, Traugott Koch, Marianne Peereboom and Andy Powell). *Metadata: an overview of current resource description practice* (deliverable for Work Package 3 of EU Telematics for Research project DESIRE No. 1004), 1996. *<URL:http://www.ukoln.ac.uk/metadata/DESIRE/overview>*

4. 'A managed information landscape' is a phrase borrowed from Denis Lynch of SilverPlatter Inc.

5. The Electronic Libraries Programme is a JISC-funded initiative. Further details about the programme and the projects it funds can be found at *<URL:http://www.ukoln.ac.uk/elib/>*.

6. Further details about the British Library Research and Innovation Centre can be found at *<URL:http://www.portico.bl.uk/ric/>*.

7. Information distributed on Mailbase list lis-elib by Rosemary Russell, Friday 15 Nov. 1996. *<URL:http://www.mailbase.ac.uk/lists/lis-elib/1996-11/0050.html>*

8. The MODELS pages are at *<URL:http://www.ukoln.ac.uk/models/>*.

9. Lorcan Dempsey and Stu Weibel. The Warwick Metadata Workshop: a

248

framework for the deployment of resource description. *D-Lib Magazine*, July/August, 1996. *<URL:http://www.ukoln.ac.uk/dlib/dlib/july96/07weibel.html>*

10. Lorcan Dempsey, Rosemary Russell and John Kirriemuir. Towards distributed library systems: Z39.50 in a European context. *Program*, vol. 30, no. 1, January 1996, pp. 1–22.

11. A. Young, P.T. Kerstein, and A. Ibbetson. Technologies to support authentication in higher education. v5. A report commissioned by the Joint Information Systems Committee of the Higher Education Funding Councils and DENI, UK. *<URL:http://ukoln.bath.ac.uk/elib/wk_papers/scoping/jisc5.html>*

12. JISC Circular 3/97 is at *<URL:http://www.niss.ac.uk/education/jisc/pub97/c3_97.html>*

13. See *<URL:http://www.ukoln.ac.uk/models/clumps.html>* for WORD and RTF versions of the profile.

Appendix – recommendations from the workshop

The actual recommendations of the workshop are listed here.

Medium term (6–9 months)

- Develop the National Agency concept (BL/JISC)
    - develop ideas in consultation with relevant players on a national and cross-sectoral scale
    - develop framework for profile
    - develop and promote the concept of national resource discovery
- Develop strategic demonstrators (eLib/UKOLN)
- Organise concertation meeting at which Z39.50 developments presented (eLib/UKOLN)
- Request eLib/SCONUL/LA to consider awareness activities for library management (eLib/UKOLN)
- Request Humanities Non-Formula Funding Committee to consider options for location of records in above context (eLib/IM)

Immediate (3 months)

- Let consultancy to develop National Agency concept (BL/JISC. Terms developed by eLib/UKOLN)
- Initiate technical profile work (eLib/UKOLN/BL)
- Initiate discussion of demonstrators (eLib/UKOLN)

Authors

Lorcan Dempsey, Director, Rosemary Russell, MODELS Project Manager, UKOLN, University of Bath, Bath BA2 7AY, UK. E-mail: l.dempsey@ukoln.ac.uk, r.russell@ukoln.ac.uk