Clumps and collection description in the information environment in the UK with particular reference to Scotland

Gordon Dunsire, Gordon Dunsire (g.dunsire@strath.ac) is Deputy Director, at the Centre for Digital Library Research, University of Strathclyde, Glasgow, UK.

George Macgregor, George Macgregor (george.macgregor@strath.ac) is a Researcher at the Centre for Digital Library Research, University of Strathclyde, Glasgow, UK.

Abstract

The role, potential and interaction of networked catalogues and collection-level description have recently been given emphasis in order that efficient resource discovery mechanisms, and the effective organisation of such resources, be facilitated within the UK's developing JISC information environment (IE). This article describes the work of CC-interop, a JISC project, and related projects that inform the development of the IE and its ability to instantiate the functional model of online resource discovery to which JISC aspires. The article reviews the evolution of Z39.50 virtual union catalogue services and collection description services that preceded CC-interop. The paper also discusses how such work is informing regional information environments, with particular reference to Scotland, and reveals how such local arrangements will benefit the wider JISC IE.

Keyword(s):

Information; Online catalogues; Information services; Scotland.

1. Introduction: the JISC information environment

The Joint Information Services Committee (JISC) in the UK works with further and higher education by providing strategic guidance, advice and opportunities to use ICT to support teaching, learning, research and administration (www.jisc.ac.uk). In particular, the JISC aims to build an online information environment that will offer convenient access to a comprehensive collection of scholarly and educational materials. An information environment “can be characterised as the set of network or online services that support publishing and use of information and learning resources. The JISC IE aims to offer the user a more seamless and less complex journey to relevant information and learning resources” (Grout, 2001, p. 3).

The functional model of the JISC information environment is being developed by the UK Office for Library Networking (UKOLN), following work it carried out on the distributed national electronic resource (DNER), the precursor of the JISC IE. The model identifies a number of stages in the process of discovering resources in an online environment (Powell and Lyon, 2003). The first stage involves the user entering the environment via a local service, which presents an
initial landscape based on a profile assigned to, or selected by, the user. The landscape is a set of collections of resources that can be accessed directly, for example Web sites, or indirectly, by using finding aids such as online library catalogues. The profile will usually limit what is included in the set, e.g. on the basis of subject, and may rank collections by significance, e.g. subject strength or accessibility. The service may present the set in text or graphical format.

In the second stage of the model, the user surveys the landscape; that is, modifies the set of collections to suit the particular task in hand. This will usually reduce the number of collections, for example by excluding those known not to be relevant, but it may also reveal new collections to add to the landscape. In the third stage, the user discovers what items of interest are contained within each collection in the landscape. This involves manual and automated searching and browsing of the metadata available from the collections. There may be a further stage, in which the user seeks further details about an item in order to identify a copy with particular attributes, e.g. a preferred location or availability profile.

The model is generally applicable to all information environments. Although developed for the JISC IE, which is UK in scope but restricted to resources for further and higher education, this paper will demonstrate that it is a useful model for regional environments which have a wider scope of resources but are restricted to smaller geographical areas.

Several of the components required to create an operational service from this model are already in place or under development. This paper describes a number of recent projects concerned with two of these: cross-searchable networked catalogues; and collection description services. Some of these projects are not yet completed, but a coherent picture is beginning to emerge of how real-world services can fit into and inter-operate within the model. This should inform the development of local catalogue and collection services to improve their integration into aggregated services, as well as assist in the planning of multi-institutional networks, collaborative collection management, and wide-area information retrieval services.

2. Background projects on clumps and collection level description

2.1 Clumps

In 1998, the third and final phase of the JISC-funded electronic libraries programme, eLib, funded the creation of four distributed union catalogue services to research and develop the use of the Z39.50 standard for wide-area information retrieval in the UK (www.ukoln.ac.uk/services/elib). This was an outcome of the third moving to distributed environments for library services (MODELS) workshop, which also anticipated the use of collection-level descriptions to present information about the catalogues and the collections described, aid navigation through the service, and help users to assess the value of pursuing access to identified resources (Dempsey and Russell, 1997). A “clump” was defined as an aggregation of catalogues, including physical union catalogues; this definition has been subsequently narrowed to refer only to distributed aggregations, and is often used even more specifically to describe aggregations based on Z39.50.
Z39.50 is a communications protocol, based on client/server architecture, which supports searching and retrieval of information in all formats in a distributed network environment (Moen, 1992). The aims of the clumps projects were to ‘‘Kick start critical mass’’ in use of Z39.50; produce model technical and other agreements to allow subsequent clumps to be justified either regionally or by subject; encouraging clumps to form; providing unifying organisation, standards and brokerage services; in the long term to see clumps extending to a truly national scale; diversity of institutions and systems; producing benefits beyond the immediate region; exit strategies – not to fund long term” (Whitelaw and Joy, 2001, p. 2).

Three of the clumps had a regional scope, and were based on existing consortia of higher education and research libraries; the fourth was focussed on a specific subject area:

1. Co-operative Academic Information Retrieval Network for Scotland (CAIRNS) included 25 members of SCURL – the Scottish Confederation of University and Research Libraries (cairns.lib.strath.ac.uk);
2. M25 link had six partners from the M25 consortium of higher education libraries in the London area (www.m25lib.ac.uk);
3. RIDING involved nine libraries from the Yorkshire and Humberside Universities Association (www.riding.ac.uk); and
4. MLO, Music Libraries Online, had as primary partners the nine conservatoire libraries in the UK, with a number of secondary partners drawn from other sectors.

Each of the projects created facilities to cross-search institutional library catalogues from a single interface. Common features of the interfaces included the ability to select which catalogues to search, and some information about the collections described by the catalogues. The projects also researched and developed a range of associated services. For example, RIDING developed reciprocal access agreements within clump members and CAIRNS used Conspectus-based subject strength measurements from SCURL libraries to implement the “dynamic clumping” or landscaping mechanism first envisaged in the CATRIONA project (Nicholson et al., 1995, p. 41).

By the end of funding in 2001, project achievements included the establishment of four working clumps, and important progress on technical Z39.50 issues. Significant developments were also made with regard to organisational elements such as collection-level descriptions and access policies. Whilst such developments stimulated a greater degree of library co-operation, the evident success of the clumps was borne out by commendable exit strategies that secured the continuation of two clumps via self funding, each representing a significant portion of the UK HE community (Whitelaw and Joy, 2001, p. 65). In fact, in addition to M25 link and RIDING continuing with self funding, CAIRNS also received assurances from SCURL of “ongoing support within the Scottish Collections Network (SCONE) project until December 2001 and then by the Centre for Digital Library Research (CDLR) until 2004” (Nicholson et al., 2000, p. 3).

Recommendation 6 of the eLib summative evaluation was that “the efforts towards co-operation and convergence within the regional and subject consortia be pursued, taking account of the non-technical developments of these projects”. The evaluation also noted that, “another important factor is whether clumps are viewed as complements to, or competitors for, union catalogue
solutions”, a thread subsequently taken up by the CC-interop project described below (Whitelaw and Joy, 2001).

2.2 Collection-level description

During 2000-2002, the research support libraries programme (RSLP) funded research into collection-level description with a number of projects (www.rslp.ac.uk). Three of the projects worked on developing services for retrieving information about collections located in, and relating to, wide geographical areas:

1. The Mapping Wales Project created collections Wales, “a bilingual online database of descriptions of research collections in Wales” (www.mappingwales.ac.uk).
2. Research and Special Collections Available Locally (RASCAL) (Northern Ireland) “consists of comprehensive descriptions of collections available to researchers in the humanities and social sciences” (www.rascal.ac.uk).
3. SCONE “provides descriptions of collections held in Scottish libraries, museums and archives, and collections about Scottish topics held elsewhere” (www.scone.strath.ac.uk).

Of these projects, SCONE was the only one to interact with a clump. One of its aims was “to investigate effective models for building and sustaining a co-ordinated Scotland-wide distributed national resource that would be conveniently accessible to researchers via the CAIRNS distributed catalogue …” (Nicholson et al., 2002, p. 5). The project developed the database of short collection descriptions used by the CAIRNS service into a comprehensive set of over 2,500 descriptions of collections held by libraries across the whole of Scotland. This enhanced database then became the driver of the “dynamic clumper” part of CAIRNS, a facility which allows the user to select catalogues by searching for characteristics of the collections they describe. Figure 1 shows an example of a collection level record in SCONE for the Glasgow digital library.

Collections can be searched by title, subject, location, and associated persons and corporate bodies (Dunsire, 2002a). The database also drives the separate SCONE collections service which provides links to local catalogues, but does not interact with CAIRNS. Further integration of the two services is being investigated in the CC-interop project.

3. The CC-interop project

In 2002 JISC granted two year funding to CC-interop (COPAC/clumps continuing technical co-operation project). COPAC is the confederation of university and research libraries OPAC, a union catalogue consisting of “the merged online catalogues of 22 of the largest university research libraries in the UK and Ireland plus the British Library” (copac.ac.uk). In particular, COPAC includes the catalogues of several libraries that are also available in the existing clumps. CC-interop “builds on the results of JISC's eLib Phase 3 programme” and “will also help pave the way for the outcomes and recommendations of the Research Support Libraries Group (RSLG)” (CC-interop project, 2002a, p. 1). At the time of writing, the project is only one year through its term.

The project comprises two work packages.
3.1 Work package A

Work package A (WPA) involves staff from InforM25 and COPAC investigating the feasibility of inter-linking between a very large physical union catalogue (COPAC) and a large virtual union catalogue (Inform25), including issues such as comparative speed of searching, de-duplication, results ranking and also comparing the accuracy both of the records themselves and the results. To date, WPA has created a copy of the InforM25 clump, previously known as M25 Link, and added COPAC as a single Z39.50 catalogue. This augmented clump was used for comparison tests of metadata retrieval from COPAC and the individual catalogues of the six institutions, which contribute to both the COPAC and InforM25 services. The report of this work confirms earlier findings of the CAIRNS project that variations in indexing practices and implementation of Z39.50 servers are two significant factors affecting the recall and precision of distributed searches (Nicolaides, 2003).

WPA is also investigating the feasibility of treating the whole InforM25 service as a single Z39.50 catalogue or server that will accept searches from remote Z39.50 clients. Like the other clumps, InforM25 currently uses a Z39.50 client to connect to multiple Z39.50 servers so that a single search created on the client can retrieve information from the different catalogues connected to the servers in the clump. If InforM25 can become a Z39.50 server itself, this would allow a Z39.50 search created from within COPAC, or indeed another clump, to be received by InforM25. The server could then pass the Z39.50 search to the InforM25 client, which would in turn forward the search to each of its individual catalogues. The ability to “clump the clumps” in this fashion is an important factor for creating a distributed union catalogue covering the whole of the UK and beyond, without duplicating the existing regional clumps.

WPA has considered the possibilities of attaching Z39.50 server software to InforM25, or developing new software for the clump that acts as both server and client, but has rejected these approaches as being too difficult to achieve within the resources of CC-interop. Instead, an open source software package, namely Java access for electronic resources (JAFER) that is being developed in conjunction with staff at Oxford University (www.jafer.org), meets the requirements and has been identified, and it is proposed that CC-interop uses JAFER and configures it to the existing M25 native Z-servers (CC-interop project, 2002b, p. 3). This work is ongoing.

3.2 Work package B

Work package B (WPB), undertaken by CAIRNS and RIDING is looking at collection level description schemas in relation to both the clumps and COPAC including issues such as target selection in clumps and developing guidelines for cataloguing and indexing practices. Music Libraries Online (MLO) has not been directly involved in CC-interop, but was to be kept informed on relevant outcomes. MLO sought funding opportunities outside of CC-interop to continue its work but these efforts have not been successful, and the clump has now been taken off-line. It is possible that specific catalogues in MLO will eventually be added to the regional clumps; for example Trinity College of Music to M25, and the Royal Scottish Academy of Music and Drama to CAIRNS. Another possibility is to develop Z39.50 clump searching facilities for music-specific collection description services such as Cecilia (www.cecilia-uk.org).
WPB has completed a report (Dunsire, 2002b) comparing the collection-level description schemata used, or in development by, the clumps, the JISC information environment, and the three regional collection description services developed by the RSLP projects. The report identified a number of data attributes to be added to the database structure to make it compatible with the full range of collection types and descriptions encompassed by the schemata. Data elements and test data for about half of the attributes have been incorporated into SCONE, and work is ongoing to identify content standards for the remainder. Corresponding amendments have been made to the Scottish collections access management portal (scone.strath.ac.uk/scamp) for updating records in the SCONE database. The report also identifies a number of issues affecting the interoperability of collection level information in general, and provides mappings to the analytical model on which SCONE and many of the schemata are based (Heaney, 2000).

WPB has created a clone of the SCONE database, populated with a representative sample of descriptions taken from the RIDING service). This is being used by RIDING to test the suitability of the augmented record structures for its collection descriptions. Copies of the user and updating interfaces have also been made in order to develop data entry guidelines suitable for both RIDING and SCONE. WPB is also working on developing and updating the recommendations from the CAIRNS project for improving the interoperability of catalogues and indexes in the CAIRNS service (CAIRNS Cataloguing and Indexing Working Group, 2000).

Future activity for WPB includes further developing SCONE as a dynamic clumping facility for cross-searching institutional catalogues in CAIRNS, RIDING, and COPAC. This work will specifically address the issue of the same metadata being available in two different clump catalogues, arising from the overlap between COPAC and the clumps. WPB will also investigate ways of exporting collection descriptions from SCONE in different formats for use by other services.

WPB is also monitoring related projects for issues arising in the area of collection description and distributed catalogue landscaping. The high-level thesaurus (HILT) project (Nicholson, 2003a) aims, in Phase II, to investigate and establish subject terminology service requirements for the JISC Information Environment, with particular reference to JISC collections and services (hilt.cdlr.strath.ac.uk). An initial draft for the service specification has been created which expects CC-interop to identify user-centred tasks for subject-related information retrieval in the distributed catalogue. The draft also indicates the desirability of adding two more attributes to collection description schemata.

The specification uses the Dewey Decimal Classification (DDC) as a common spine for mapping different subject terminologies. When a particular subject term is matched against the sets of terminologies within the server, and subsequently disambiguated by the user, the DDC numbers associated with the term can be used to select appropriate collections for item-level searching, provided the collection description schema accommodates DDC-based collection strengths; that is, a range of classification numbers for general collections, and specific numbers for subject-specific collections. For example, the term “lotus” may be disambiguated to the category of flowering plants, and the corresponding DDC number used to identify collections with strengths in the subject area of botany.
The terminologies server will also have information about which particular terminology sets the term is associated with. If the server can identify the specific vocabulary used for item-level description within a collection, for example library of congress subject headings, the appropriate term for that set, for example “lotus”, can be used manually or machine-to-machine to carry out item-level retrieval on the collection (Dunsire, 2003). This suggests that the collection description schema should also accommodate information about the terminology set used for item-level description of the collection, whether based on topical headings or classification schedules.

These preliminary results from CC-interop and HILT allow us to see how specific services for collection-level description, terminology resolution, and multi-institutional information retrieval might interact within the functional model, as shown in Figure 2, of an information environment. We can further propose a broad operational model for a regional information environment for Scotland, incorporating the results of specific Scottish activities described below.

The Figure shows how the terminologies server based on the HILT model and SCONE services interact with each other and initial entry landscapes such as Scottish cultural portal referred to below, in order to facilitate the entry and survey functions. SCONE will create its own initial landscapes for CAIRNS and RIDING users during the CC-interop project. This project will also integrate access to the Scottish collections described in COPAC, while distinguishing between duplicate sets of metadata. The figure is a simplification, and refers only to projects mentioned in this paper.

4. Other related developments in Scotland and elsewhere

4.1 HaIRST

CAIRNS and SCONE are also being developed as components of a number of other projects. The harvesting institutional resources in Scotland testbed project (HaIRST) is part of the focus on access to institutional resources programme (FAIR) funded by JISC. The project is investigating the use of the Open Archives Initiative (OAI) protocol for harvesting metadata from three universities and ten further education colleges in Scotland. The project's deliverables include two-way metadata mappings to support further discovery and disclosure to and from CAIRNS and collection-level metadata databases such as SCONE (hairst.cdlr.strath.ac.uk). The intention is to harvest metadata about resources created in the institutions, ranging from e-prints to teaching materials, and create an online union catalogue that will become part of CAIRNS.

Although the universities are already members of CAIRNS, none of them is also a member of COPAC. However, if the HaIRST approach is subsequently used by a COPAC member, the situation will arise where the metadata of some collections will be duplicated in at least three CAIRNS catalogues: the local library catalogue, the HaIRST catalogue, and COPAC. The secondary harvesting of HaIRST metadata by other services will also engender similar duplication. Appropriate collection metadata will be required to allow coherent landscaping of these different views of the information environment.
4.2 SPEIR

The Scottish Portals for Education, Information and Research (SPEIR) project (Nicholson and MacGregor, 2003) has recently been funded by the Scottish Library and Information Council (SLIC) to “research into distributed information infrastructure requirements for the two year Scottish cultural portal pilot, and the one year public library CAIRNS integration proposal, and, where appropriate, to develop associated pilot facilities” (Nicholson, 2003b, p. 1). Many project objectives involve SCONE and CAIRNS, including seamless integration of the cultural portal pilot into CAIRNS and related initiatives such as the SCONE collections facility and extending the CAIRNS pilot to further education libraries and public libraries. The portal pilot is creating an information environment for Scottish cultural resources, and includes its own catalogue of online resources. It will use SCONE and CAIRNS as landscaping components. Funding is also being provided to assist public libraries in adding their resources to the environment by implementing Z39.50 servers and joining CAIRNS, or employing other techniques such as metadata harvesting. SPEIR also aims to build on the work of HILT to specify terminologies requirements for the Scottish cultural portal pilot, the people's network in Scotland, and other purposes.

4.3 London-based projects

In addition to these developments in the information environment of Scotland, several projects involving clumps and collection description services have created a similar focus of activity in the London area including:

- the Find it in London Project (2002) which “sought to demonstrate that it was possible to create a Web-based look up tool covering collection level descriptions at a high level, from all sectors and domains”;
- archives in London and the M25 area (AIM25), is a “project to provide electronic access to collection level descriptions of the archives of over 50 higher education institutions and learned societies within the greater London area.” (www.aim25.ac.uk); and
- what's in London's libraries (WiLL), is a project which “aims to link up all the catalogues and community information databases of London's public libraries … most of the searching will use the Z39.50 protocol …” (www.llda.org.uk/will).

5. Conclusions

Services arising from the initiatives described in the previous section are, in the main, being created as separate entities, specifically defined by sector and domain. A more general information environment for London, and other regions, which integrates multiple clump and collection services will become possible if key deliverables of the CC-interop project are met. The ability to pass searches from one clump to another is as useful in creating regional catalogue networks from smaller components as creating national networks from regional ones. The use of interoperable collection description services will enhance the information environment by improving coherency and flexibility, giving the user greater control and better tools for surveying the landscape.
In Scotland, however, projects have tended to extend and expand CAIRNS and SCONE rather than create separate services. This allows us to fit the existing and planned services described in this paper into the functional model of the JISC information environment, as illustrated in Figure 2.

The CAIRNS project report referred to “An embryonic Scottish national networked information service or Scottish Portal that continues to develop and offers the potential of integrated access to all publicly available research materials, learning and teaching resources, and public information services in Scotland, whether digital or non-digital …” (Nicholson, 2000, p. 3). The subtitle of that report poses the question: an embryonic cross-sectoral, cross-domain national networked information service for Scotland? CC-interop and the other projects related to CAIRNS can allow us to respond, three years later, with an emphatic yes.

![Figure 1A collection-level description displayed in the SCONE service](image-url)
**Figure 2** Operational model of the Scottish information environment

**References**


