68th IFLA Council and General Conference August 18-24, 2002

Code Number: 022-144-E
Division Number: IV
Professional Group: Classification and Indexing with Cataloguing - Workshop
Joint Meeting with: CILIP Cataloguing and Indexing Group
Meeting Number: 144
Simultaneous Interpretation: -

Joined up indexes: interoperability issues in Z39.50 networks

Gordon Dunsire
Napier University Leaning Information Services
Edinburg, UK

Abstract:

The paper discusses issues in the interoperability of indexes to metadata records in distributed information retrieval networks, based on the findings of the CAIRNS and SCONE projects. The Co-operative Academic Information Retrieval Network for Scotland and Scottish Collections Network Extension projects have evolved into embryonic services which fit together to provide user-driven collection identification and selection mechanisms and the ability to cross-search related metadata for item discovery and access. The CAIRNS Cataloguing Issues Working Group identified a number of factors affecting cross-searching of metadata indexes for authors, titles, subjects and control numbers, including local cataloguing policies, content standards, and index structures. The SCONE project has identified issues in subject indexing at the collection level, in particular the relationship between collections with specific subject content and general collections for which Conspectus-type subject strength mappings are appropriate. The papers discusses these findings in a cross-domain context.

The Co-operative Academic Information Retrieval Network for Scotland (CAIRNS)[1] and Scottish Network Extension (SCONE)[2] projects have resulted in two embryonic national information retrieval services for Scotland. The CAIRNS service[3] is a "one-stop-shop" for cross-searching some 20 online catalogues, including those of most of the Scottish universities and the National Library of Scotland. It employs Z39.50 for broadcast searching, with a web-based interface that allows users to save time and reduce information overload from irrelevant catalogues by selecting a sub-set of catalogues before commencing the search. Selection, known as "dynamic clumping", can be made directly on a list of

catalogues, or by using information stored in the collection level description record associated with the catalogue. This includes the subject strength, geographical location, and special sub-collections of the collection which the catalogue describes. The collection level descriptions are maintained by the SCONE service[4], which currently contains records for 3,500 collections and sub-collections, and their associated catalogues. SCONE and CAIRNS are semi-integrated, and are expanding to cover libraries in all sectors and domains; further integration will be carried out as part of several new research projects including phase II of the High Level Thesaurus (HILT) project[5] and the COPAC/Clumps Continuing Technical Co-operation project. Frequently used combinations of catalogues and indexes to search are made available in CAIRNS as static "mini-clumps" which can be invoked as shortcuts through the selection process. An example is the "Napier Health Reclassification via ISBN" mini-clump which selects four catalogues and the ISBN search, used by paraprofessional staff at Napier University to identify Dewey Decimal Classification numbers for materials on health subjects.

A significant activity carried out during the CAIRNS project was the establishment of a Cataloguing Issues Working Group (CCIWG), consisting of representatives from the CAIRNS member libraries. This group met several times during the project to discuss cataloguing policies and practices which might affect the interoperability of cross-searching; the group is now being reactivated as part of the Confederation of Scottish Mini-Clumps (CoSMiC)[6], an umbrella organization for disseminating information about distributed searching and discussing associated issues. The CCIWG produced a set of recommendations for improving interoperability by changing local practices in cataloguing and indexing, and retroconverting legacy data, published as Appendix D of the CAIRNS Final Report[7]. These recommendations were subsequently adopted as policy by the Scottish Confederation of University and Research Libraries (SCURL)[8]. It should be noted that at the beginning many cataloguers not entirely familiar with their local policies and practices, or why they had been developed. Participation in the group forced a re-examination of the local environment, and in some instances an immediate revision of out-of-date practices. An important outcome of the work of the CCIWG was to make cataloguers aware of the impact of local conditions on the effectiveness of distributed information retrieval networks; "think globally before acting locally" has become something of a mantra in Scottish libraries.

The CCIWG focussed on the standard CAIRNS searches: Author, Title, Subject, ISBN and ISSN. The "simple" general keyword search was added to CAIRNS at a later date, and was not discussed. Some common themes affecting the consistency of all types of searches were identified:

- The mapping of metadata record elements to the index; what can be searched in the index.
- The format of index entries and availability of search modes; how the index can be searched.
- The depth or granularity of metadata records; precision and recall in searching the index.

Libraries supplied a detailed mapping of metadata record elements to each of the index types, in terms of MARC tags and subfields, which could be compared to the MODELS Library Interoperability Profile[9], now superseded by the Bath Profile[10]. Divergence was found for every index, with some libraries mapping elements not included in the MODELS set, or not mapping elements recommended by MODELS. Comparison tables were published as part of Appendix F of the CAIRNS Final Report[7]. For example, Tables F5 and F6 cover title indexes and indicate that many catalogues do not include title elements from author-title added entries or author series statements. The least divergence was found in ISBN and ISSN indexes, as shown in Tables F9 and F10, as might be expected from the small number of MARC tags that can contain bibliographic control number data; even so, several libraries only offer a combined ISBN and ISSN index, and some include the contents of the record control number tag 001 which may, or may not, contain an ISBN or ISSN. The lack of consistency in the mappings across the CAIRNS libraries means that inconsistent results are returned from multi-catalogue searches. For example, a search for a title by a different author in a work bearing a collective title will fail if author-title

entries are not indexed, even though the work may be held by the library in question. While local users of that library may know about this aspect of indexing policy, it is likely that external users will assume the work is not in stock. They may not assume that their search strategy is incorrect because they will retrieve the metadata for the work held in another library. The Group recommended that libraries develop a common mapping for each index type, and ensure the availability of indexes required for conformance to the Bath Profile.

The format of index entries raises two general issues: form and completeness. The content of author and control number indexes is usually created in normalized form. For personal authors, this means presenting the family name first, with given names and other information following; for control numbers this involves stripping out punctuation such as hyphens and blank spaces. Compound family names create ambiguity in what constitutes the actual family name, and it is possible for "Van Winkel" to be indexed under "Van" in one index, and "Winkel" in another, with obvious problems for the user of an author browse index. Leaving blank spaces in ISBNs can cause an exact match search to fail. Completeness of index entries primarily affects author searches. Libraries may follow standard rules for distinguishing between similar names for different persons or corporate bodies, say by adding initials, full given names, and dates successively until the names are distinguishable, but this is often done only in the local context, relative to the local catalogue. A problem then arises when cross-searching that catalogue with another, for the distinction between different authors may be lost in the wider context. The Group recommended the use of the fullest form of names, using a common authority file, as a solution. Completeness may also be an issue with titles if the local system operates a stop-word policy. While stopping common articles and conjunctions such as "the" or "and" has a trivial effect on searching, some libraries have policies of stopping frequently occurring words such as "Scotland". This is often because of legacy system infrastructure restrictions, and is expected to disappear as systems are modernised. In the meantime, this will result in inconsistent retrieval during title keyword searches.

CAIRNS offers two modes of searching indexes: "standard" and keyword. Standard search mode is for matching search terms to the beginning of index entries, and in some cases the whole of the index entry. Keyword searches usually match terms against whole words found anywhere in the index entry. Problems arise when a particular mode is not supported, for example "author keyword", by the local cataloguing system, or has not been implemented. Solutions lie in the conformance of local vendor systems to the Bath Profile, or by adding the appropriate index; cataloguing policies have little impact.

The Group spent a surprising amount of time on ISBN and ISSN indexes. In particular, the discussion raised the issue of the depth of cataloguing, or the granularity of metadata records, in the context of multi-part works and serial formats. Some libraries create a single record for a multi-part work, with parts being described in notes and indexed using added entries, whereas others create separate records for each part. In the latter approach, some libraries link the records explicitly as analytics, while others rely on implicit linkage using added title entries. There was a similar divergence in the way different serial formats are handled, with some libraries creating a single record for both print and electronic versions, and others using separate records. This can have a serious negative impact on control number searches. An ISBN search can retrieve metadata for the whole set or individual part, forcing the user to examine the metadata in detail, potentially at the item or copy level, to ensure that the set or part is actually held by a particular library. An ISSN search for an electronic serial can retrieve metadata for a print format irrelevant to the user's needs. Similar problems were identified for monographic series, which are variously treated as serials, single monographs, or multi-part sets. Although control number searches are affected much less by other issues, they are extremely important to users for identifying works uniquely. In particular, information professionals rely on them for collaborative collection management, and for de-duplication in union catalogues. The CCIWG recommended stricter adherence to international cataloguing standards to alleviate the problem, for example treating electronic and print versions of a serial as different works and ensuring that ISBNs are recorded with appropriate qualifiers. It should be noted that although the

proportion of items that can be identified by ISBN and ISSN in networked information retrieval is decreasing as online archive and museum finding aids become available, and as increasing numbers of electronic resources without standard numbers are catalogued by libraries, this may not significantly impact on user needs for comparing like with like. Archive and museum resources tend to be unique, requiring only a single copy of metadata, so de-duplication is not an issue. Multiple copies of metadata for electronic resources may be reduced if cataloguing practices are changed to resolve other problems with online resources.

The Group noted that, while granularity issues for print resources were largely confined to the areas already mentioned, the situation for web-based and other electronic resources was potentially much more of a problem. Some CAIRNS catalogues contain metadata for electronic resources which are deeply embedded in web sites, without reference to other components of the web site. For example, the SLAINTE catalogue records individual poems which are part of a collection of digitised poems which in turn are part of a web site dealing with literature; the "collecting policy" of SLAINTE covers certain writers and poets only, so other poems are not recorded. It is clearly possible that another catalogue records the collection of poems as a single entity, and yet another may record the web site itself as a single entity. Presenting the results of a cross-search across all three catalogues in a coherent, consistent and complete way to the user is something of a challenge, even supposing that a single search could retrieve all three records. Although the Group did not make any recomendations in this area, the CAIRNS project suggested a possible solution, described in Annexe B of Appendix B of the CAIRNS Final Report[7]. The proposal suggests that the CAIRNS network needs only to record web resources once, for use by the network as a whole, if there are no local restrictions or requirements for access. If such records form a separate catalogue within the system, they can be automatically cross-searched by treating the catalogue as pre-selected in any dynamic clump or static mini-clump. Techniques of explicit bibliographic linkage can then be applied to the single catalogue to improve the coherency and consistency of searches. Some work has been carried out by the Centre for Digital Library Research (CDLR)[11] which maintains CAIRNS and SCONE, and the Scottish Library and Information Council (SLIC)[12], to further this proposal by assessing the potential of the OCLC CORC service[13].

The CCIWG had little to recommend about subject indexes and searching. Although Library of Congress Subject Headings is the most common scheme in CAIRNS libraries, it is only used in half of them, as shown in Table F16 of Appendix F of the Final Report[7]. The situation is similar for classification schemes, and although Dewey Decimal Classification is most prevalent, multiple editions of it are in use, with significant differences between editions. The Group recommended the adoption of a single subject authority scheme; the brevity of the recommendation is indicative of the Group's view of the likelihood of this happening in reality.

Further issues in subject retrieval were identified by the HILT project. A Focus Group of representatives of museums, archives and libraries produced a report[14] that identified the desirability of cross-searching online catalogues and other finding aids by subject, along with issues that restrict the possibility of achieving this. Issues were found to be similar to those identified by the CCIWG: resourcing, legacy material, disagreement about standards, and differing standards in use. A subsequent workshop, again with cross-domain representation, showed a clear consensus that the favoured approach to improving cross-searching by subject was to set up a pilot service to map subject terminologies between the major schemes in use[15]. Funding has been secured to do this.

In Scotland, there are separate services for cross-searching online finding aids within each domain: CAIRNS and SCONE mainly cover libraries; the Scottish Cultural Resources Access Network (SCRAN)[16] is primarily concerned with museum resources; and the Scottish Archives Network (SCAN)[17] provides access to archival collections. Each service uses different approaches to subject retrieval; as yet, no attempt has been made to cross-search between the services. However, work on integrating CAIRNS and

SCONE has uncovered a particular issue that is likely to affect subject retrieval across domains. CAIRNS uses Research Collections Online[18], a three-tier thesaurus scheme based on Conspectus, to indicate the subject strengths of general collections. A standard set of subject headings, each with a Conspectus level, is attached to the relevant collection level description; headings can be searched to identify which collections have strength in that subject. The structure of this approach is horizontal: a fixed set of terms with shallow granularity exhaustively applied to general collections. SCONE collection descriptions also have subject headings attached, but only those headings which are relevant to the specific subject of the collection, and with no indication of the "strength". This approach is vertical: a dynamic set of terms with deep granularity relatively applied to collections on specific subjects. If the two services are to be fully integrated, a means of providing users with a coherent facility to identify collections which are either "about" or have "strength" in a particular subject needs to be developed. The challenge to provide better interoperability for access to library, archive and museum resources, particularly on a subject basis, has a long way to go.

References and links

(1) CAIRNS project: http://cairns.lib.gla.ac.uk/ (checked 25 Apr 2002)
(2) SCONE project: http://scone.strath.ac.uk/ (checked 25 Apr 2002)
(3) CAIRNS service: http://cairns.lib.strath.ac.uk/ (checked 25 Apr 2002)
(4) SCONE service: http://scone.strath.ac.uk/service/index.cfm (checked 25 Apr 2002)
(5) HILT project: http://hilt.cdlr.strath.ac.uk/ (checked 25 Apr 2002)
(6) CoSMiC: http://cosmic.cdlr.strath.ac.uk/ (checked 25 Apr 2002)
(7) CAIRNS Final Report: http://cairns.lib.gla.ac.uk/cairnsfinal.pdf (checked 25 Apr 2002)
(8) SCURL: http://scurl.ac.uk/ (checked 25 Apr 2002)
(9) MODELS Library Interoperability Profile:
http://www.ukoln.ac.uk/dlis/models/clumps/technical/zprofile/zprofile.htm (checked 25 Apr 2002)
(10) Bath Profile: http://www.ukoln.ac.uk/interop-focus/bath/1.1/ (checked 25 Apr 2002)
(11) CDLR: http://cdlr.strath.ac.uk/ (checked 25 Apr 2002)
(12) SLIC: http://slainte.org.uk/slic/slichome.htm (checked 25 Apr 2002)
(13) CORC: http://www.oclc.org/corc/ (checked 25 Apr 2002)
(14) HILT Focus Group Report: http://hilt.cdlr.strath.ac.uk/reports/focus2603.html (checked 25 Apr 2002)
(15) HILT Workshop: Report and Conclusions:
http://hilt.cdlr.strath.ac.uk/dissemination/workshopnew.html (checked 25 Apr 2002)
(16) SCRAN: http://www.scran.ac.uk/ (checked 25 Apr 2002)
(17) SCAN: http://www.scan.org.uk/ (checked 25 Apr 2002)
(18) RCO: http://scurl.ac.uk/vuc/rco.html (checked 25 Apr 2002)