Lumps in the Clump: More Stirring Required?

Gordon Dunsire

A clump is defined as “two or more autonomous information retrieval systems networked to encourage wider access and foster collaborative collection management”. This is a broad definition; many of the issues are relevant to any service linking separate information retrieval systems together, irrespective of technical infrastructure or the specific purposes of the service. Beyond simple hard-wiring, integration of service elements from separate systems requires a degree of interoperability - the ability to exchange and format information from one system to the other.

A lump is then defined as any barrier to such interoperability. Behind any such barrier, the service may retain features of the originating system which are incompatible with the “clumped” service.

Z39.50

Z39.50 technology allows the interoperability of search facilities and integrated display of results in a clump. It allows the formulation of a search in a local system (known as a Z client), and translates the search elements into a standard, neutral protocol which is sent to another system (the Z target). The search is understood by that remote catalogue. Search results are treated in a similar way, formatted and delivered to the enquirer' s local system.

Although this standard has been available for several years, there was little real-world implementation of it to begin with. This may have been due to technical factors. Unwillingness to share resources or reduce local efficiencies to improve general effectiveness - the collaborative cringe - were also, and remain, significant problems. Phase 3 of eLib (the Electronic Libraries Programme), provided funding in 1999 to set up experimental Z39.50 clumps to explore practical issues connected with their implementation, development, and use. All of these clumps are centred on HE libraries, although they may include non-HE systems. They comprise the

CAIRNS, Riding and M25 projects, which have general academic coverage, and Music Libraries Online which is subject-specific.

At around the same time, implementers began to discuss the need for ways of measuring the conformance of individual systems to the Z39.50 standard, to encourage system developers and suppliers to follow it. The resulting Bath Profile is a document specifying a number of levels of conformance to Z39.50. There is evidence that this is being adopted by system vendors. It remains the case, however, that none of the eLib clumps currently meet even the lowest level of conformance specified in the Profile.

CAIRNS

CAIRNS (the Co-operative Academic Information Retrieval Network for Scotland), is one of these clumps. Based in Scotland, it links together the online catalogues of members of SCURL (the Scottish Confederation of University and Research Libraries), which includes all of the universities, the National Library of Scotland, and the two largest public library services.

The public libraries, and a couple of the Universities, do not have the necessary Z39.50 capability and cannot be searched via CAIRNS. However, all member libraries took part in discussions about factors affecting the interoperability of the clump, particularly cataloguing and indexing policies.

Lots of lumps were identified during the lifetime of the original project. Many of the barriers to interoperability are familiar from past experience in creating union catalogues, and methods for reducing or removing them are well-understood.

One of the newer lessons to be learned from CAIRNS is that changes to the local information retrieval system can make interoperability worse, and sometimes non-existent. Z39.50 systems require both a target server and a client. Local changes made to the server set-up can render any of the clients inoperable against that server. Such changes are not confined to parameters specific to Z39.50. Access to indexes can be affected by how they are implemented, and a search which assumes truncation of the search term will produce misleading results when carried out on an index which does not support truncation. Changes to the server record syntax may impair the display of retrieved records by the client system, possibly dropping important content. The Bath Profile specifies index structures and record syntaxes, so system administrators have a set of guidelines for a standard Z39.50 server, but compliance is difficult in practice at the moment. For example, the Profile does not specify UKMARC, which is the most widely used in Britain.

Any changes made to a Z target set-up must be communicated to actual and potential Z client administrators, so that the client set-up remains synchronised with the target. System administrators need to think globally while acting locally.

The Cairns project identified several ways of improving communication about Z server implementations. At the very least, systems administrators should publish the set-up details on a local website, in an easy-to-find place. If a Z client fails to interoperate properly with the server, the client's system administrator can check the details and determine what amendments are required to the client set-up. This approach assumes that system failure will alert staff to the problem. To prevent failure occurring, changes to the server set-up must be broadcast immediately to client administrators, perhaps via an email list. This in turn raises issues of list maintenance. In particular, "amateur" administrators, with their own personal clients, are unlikely to be members of the list.

Within a clump, server administrators should be able to update a set-up record held centrally for use by the service. This can take effect immediately, and is under the control of the local admin istrator. The CAIRNS service will implement such a facility as part of a future SCONE (Scottish Collections Network) staff portal. To sustain maximum interoperability in these circumstances, the clump should also include a proxy service and act as a Z server itself, and be promoted as the standard portal for information retrieval in its area of operation.

The CAIRNS Clumper

The CAIRNS clumper, the mechanism driving the service interface, is implemented entirely in software, and consists of three functional components. First, a public web interface provides an input screen for user-generated searches, and display screens for search results. This is maintained using Macromedia Ultradev and ColdFusion technologies. Second, a SQL database holds records containing Z server set-up information for each member of the clump. This data drives some of the options presented to the user input screen. Finally, a Z client capable of broadcast searching takes search options and technical data to search a number of the CAIRNS Z targets simultaneously. This client is part of the Horizon system developed by epixtech, inc.

This configuration has a number of scaling problems. As each Z target is added to the system, the bandwidth required to search all targets increases, along with the response time. A search carried out on all targets is likely to increase the number of results returned to the user, which can be detrimental in many circumstances.

Some of the scaling problems can be resolved by allowing the user to select a sub-set of the Z servers available in the clump for searching. The CAIRNS service offers such a “dynamic clumper”. It allows selection of Z targets by location (strictly speaking, the place where the items described by the target catalogue are stored), and by subject. The SCURL RCO (Research Collections Online) schema, based on Conspectus, gives subject strengths of the main general collections. The SCONE project is subject indexing special collections.

Miniclumps

The data required by the dynamic clumper is stored with the technical Z target data in an extended database based on the RSLP (Research Support Libraries Programme) collection descriptions schema. One advantage of this approach is that selected sub-sets of the CAIRNS targets can be stored as pre-set choices for special groups of user. This encourages selective use of bandwidth, as well as providing a short-cut through the dynamic clumper. In CAIRNS, such pre-set sub-sets of targets are known as "miniclumps". The CAIRNS project identified three types of miniclump; regional, subject, and special.

Regional library partnership groups in Scotland all aim to share resources and encourage wider access; many see the availability of an integrated, networked union catalogue as a key facility. Nearly all such groups contain at least one active CAIRNS member.

Miniclumps based on subject strength can make searches in that subject area more efficient and effective. For example, academic staff at Glasgow University have identified a sub-set of CAIRNS targets which they consider to be the most useful for checking history resources.

Special miniclumps, like the temporary miniclump created to assist a reclassification project at Napier University, can be set up for use by library staff. CAIRNS targets likely to contain appropriate Dewey Decimal classification numbers for the stock in question can be selected. This miniclump also used a sub-set of available searches, as only ISBN and title searches were required.

Significantly, all "real-world" miniclumps identified by CAIRNS are actually, or potentially, cross-sectoral and cross-domain. However, outside the CAIRNS members themselves, the take-up of Z technologies remains low. There is no doubt that the CAIRNS project encouraged many HE libraries to install Z servers, but extending the CAIRNS service beyond this sector requires a much greater involvement of FE, public and schools libraries, as well as archives and museums.

The CAIRNS project identified a number of reasons for the low density of Z targets outside HE. Some smaller system vendors do not supply a Z server; others charge a cost which is excessive if the focus is purely on local services. Many organisations lack adequate levels of technical support. Some Z target software simply does not work properly, or is incompatible with other software in the local environment. Institutional policies on data security can prevent the installation of any software which involves data flowing across network firewalls. And active collaboration remains, for many, fine words on paper, with little financial encouragement to make it work effectively.

Clump as Community

The expansion of clumps can be improved if members cooperate and collaborate as a community. Those with the necessary expertise and experience can share it. Collective adoption of standards is more effective than piece-meal development. Economic and political problems may be better addressed if there is a louder voice expressing the interests of clump members.

The architecture of the CAIRNS dynamic clumper allows miniclumps their own distinctive look and feel, but based on the same set of underlying control records. This community does not just involve system administrators and library managers. CAIRNS identified lumps at all levels of the retrieval process, in areas involving all professional roles. As well as interoperability problems within the set-up and maintenance of Z technologies, barriers can be found in many other areas of bibliographic retrieval, and dealing with them is equally important.

The format and content of indexes affects interoperability. Assumptions about "normalised" data structures such as form of personal name, dates and language codes have to be examined carefully. They may be different for different sectors and information domains. Semantic variability is another problem; searches for "porridge" and "porage" will produce quite different results unless there is a mapping (using a thesaurus or subject authority file) from one form to the other, either at the target or clumper end.

Local cataloguing policies determine which items of stock have searchable metadata, and the depth of description, which will vary. In general, all "granularity" in cataloguing practices and procedures will cause lumps, but not all lumpiness is due to granularity!

The Cosmic Scone

The Centre for Digital Library Research has inaugurated CoSMiC, the Confederation of Scottish Miniclumps. Initial work has focussed on sharing expertise gained in the CAIRNS project with prospective regional miniclumps, and in providing a forum for the exchange of views. Not all groups will create their own Z39.50 clumps, but many of the interoperability issues are common to any technical solution to networking catalogues, and it is important that the broader definition of "clump" is kept in mind.

SCONE, the Scottish Collections Network Extension project, builds on the infrastructure created for the CAIRNS service. It provides a means of adding non-Z 39.50 targets to the Scottish clump, albeit with a high degree of lumpiness. These additional catalogues and other finding aids can be selected using standard dynamic clumping facilities, but instead of being searched in a single-search broadcast mode, they can only be searched by direct connection and use of the local interface. SCONE covers all library sectors, as well as some museum and archives collections, and has the potential to provide a user-driven, dynamic information landscape for the whole of Scotland.

In spite of the problems, these projects work. They can be improved if all constituent organizations and individuals develop their local policies for the benefit of all. More stirring is definitely required. The good news is that even a little effort breaks the odd lump or two.

This is the edited text of a presentation at UmbrelLa 6, 5-7 July 2001, Manchester Conference Centre

The PowerPoint Show format slide presentation can be downloaded from http://catriona.napier.ac.uk/docs/lis/dunsire/lumps.pps. A suitable plug-in is required to view this format.

The author can be contacted at g.dunsire@napier.ac.uk

The CAIRNS Project website is at http://scone.strath.ac.uk