



Sketch for an automated approach to
'scoping ahead' in Digital Scotland
and the DNER based on distributed
'collection strength' indices

Final Report of the RSLP SCONE project
Annexe A.4

Note: This sketch aims to show that this is an area worthy of further research. There is no suggestion that all questions associated with this area are answered in this annexe.

The CURL iCAS study Evaluator's Report¹ concluded that:

*The automated approach [to measuring collection strength] has great potential, offering the long-term possibility of objective measurement, marginal costs, support for deep resource sharing, immediacy of collection 'strength' data, and a much improved service to users, but it also entails major implications in respect of what will be required of institutions before this potential can be fully realised. It should, for example, be possible one day for staff and users to have access to a 'clump' of cross-searchable collection strength indices based at COPAC, the National Libraries, and members of CAIRNS, M25 and RIDING. These indices could be built automatically at marginal cost by library management systems operating in real time as institutional cataloguers added records, could offer up to the minute data to both users seeking materials and staff engaged in deep resource sharing, and could theoretically allow filtering for uniqueness (regardless of numerical strength), language, research or teaching code, location, access category and other relevant elements. As with other automated approaches, however, full functionality will only be possible if institutions are willing to undertake significant long-term work on metadata, if common classification and subject schemes are adopted, if institutions will add additional collection strength data to item records, if they will pay for the creation and maintenance of associated indices, and so on. Exactly how the distributed index described above would be designed, and what the specific implications for institutions would be requires more detailed work as described below under 'Recommendations'. **There may be value in considering examining this approach in the arena of e-resources in the first instance, there being less of a problem with legacy metadata in this area.***

This Annexe (A.4) explores this latter idea further, focusing in particular on the idea of utilising automatically produced collection strength indices to allow users to 'scope ahead' in distributed digital collections such as the DNER or Digital Scotland or NOF, utilising an extension of the ideas of the CAIRNS dynamic clumper and drawing in the significance of certain aspects of HILT Phase II to the equation.

The assumption behind this sketch is that the future of information systems at every level, from Global, to UK, to Scotland wide, to Glasgow-wide is a distributed one. This does not mean that large physical union catalogues like COPAC will disappear, only that complete coverage of all of the interesting learning, research, informational, cultural, and recreational materials will require intelligent access to a wide range of geographically distributed repositories of digital and non-digital materials, rather than to one single central system. In this context, intelligent access means that users are offered a mechanism similar to what CAIRNS calls dynamic clumping to enable them to select a manageable sub-set of all of the available repositories relevant to their needs at any given point in time.

¹ Written by the SCONE Project Director

Once this subset is selected, users will have a number of options, depending on circumstances, need, and the search facilities available:

- For hard copy collection with no online catalogue, transcription of access instructions (e.g. which bus route to use) and a visit to each relevant file.
- For collections with some form of stand alone online catalogue, an online search of each relevant catalogue, followed as either visits, ILL or DD requests, or for e-resources electronic delivery.
- For collections whose catalogues are interoperable, a cross search of the relevant catalogues, followed by visits etc.
- For collections where a cross-search reduces the available search facilities, perhaps a follow up stand alone search, followed by visits etc.

If the dynamic landscaping mechanism has 100% efficiency and users are exhaustive in their subsequent searches, at least until they have the resources they seek, user queries will be answered to the extent feasible in the available distributed repositories. The question addressed in this sketch is now to best optimise the efficiency of the landscaping mechanism, both in the sense of maximising reliable user navigation of the internet in the distributed environment and in the sense of minimising the cost of effort required to provide it. This question is at the core of the SCONE research question addressed in this annexe - the requirement to identify a more objective, less labour intensive means of measuring collection strength than the Conspectus based methodology that lies at the heart of the current CAIRNS dynamical clumper.

The current mechanism is based on the subject strengths of collections as determined by library staff utilising a conspectus based methodology. Users can either search for a subject heading that reflects their research requirement or browse down a hierarchy to find it. At the end of the process they find a sub-clump of CAIRNS catalogues logged as having either a teaching level, a research level, or a comprehensive level strength in the subject concerned and can then conduct a cross-search of the subset. Discussed extensions of this in CAIRNS include landscaping on user type (e.g. collections relevant to undergraduates), on sector or domain (HE or FE collections only), geographical region (e.g. just Scotland, just London), language, or user task. Although it has a number of recognised deficiencies, concerned in particular with the reliability and usefulness of the subject strength data, its currency, and the labour intensity of the method of compilation, it is true, and not entirely trivial, to say that the mechanism works. Given reliable, useful, and current data, collected at marginal cost, it would be possible even in a distributed system encompassing a very large number of repositories, to reduce the number of repositories of likely interest to the user to a manageable proposition.

Typically, the user will either come to the proposed extended landscaping mechanism with a good deal of information of his or her requirements as regards likely repositories or, if not, will be able to acquire it through an explanatory session with a suitably informed user help facility. The system will prompt the user for data about the information requirement that will, given the existence in the system of a database of appropriate collection level information about the repositories available, enable the system to landscape a subset of repositories for the user to investigate further through one or other or a mix of the methods listed above.

This information might include:

- Data of the subject of the search chosen from a drop down list, preferably at the lowest appropriate level of granularity.
- Data on the task to be carried out (e.g. all graduate level books on the subject, all research papers on the subject).
- Geographical data (e.g. initial search in this geographical area only (where I will have access)).

- Sectoral data (e.g. HE catalogues or FE catalogues only).
- Domain data (e.g. libraries only, archives only, museum and libraries only).
- Data on format (e.g. electronic only).
- Data on cost (e.g. free only or pay by credit card).
- Data on language.

It might also include data available to the system about the user. Current educational level in this subject, resources available free, past preferences, and so on. By mapping this data against the database of collection level descriptions the system should be able to offer the user a limited number of probable repositories to search and probably even rank them in order of probable usefulness.

Let's say in a world-wide system this enables users to narrow their choices down to 100, or even 200 repositories. This is still a manageable number, even if it is considered necessary to limit cross searches to, say, 15-20 at a time. If the user is likely to be satisfied with a few useful documents, he or she can stop after these have been found - and a researcher looking to be more exhaustive will presumably expect to spend longer on the task in hand. Moreover, where the scope of the distributed system is more limited - in the DNER, say, or Digital Scotland or all NOF projects - the likely number of repositories will be lower and even more manageable. In short, provided we can gather reliable, useful, and current collection level information at marginal cost to feed the landscaping database, this approach can work - which is to say, it can make intelligent navigation in a large distributed information system a feasible proposition.

Cross-searchable collection strength based indexes built by local systems at the distributed repositories themselves as described in Annexe A.3 offer a possible means of achieving this - a means that should be implementable for either small or new e-resource collections and simulated (with some limitations) for larger legacy systems in the short term and, given the will and the resources, implemented in the longer term for even this latter group. Although there maybe exceptions, small or new repositories are most likely to be repositories of electronic resources, so the main focus here is on collections of this type. However, small or new hardcopy repositories could also be included in the research proposed at the end of this annexe should appropriate candidates exist.

Recognised problems with the current CAIRNS mechanism include:

- a) A lack of objectivity and consistency in the measurement of what constitutes 'strength' in a particular subject, so that the navigational information is less relevant and useful than is ideal in such a system.
- b) Staff time and effort entailed, and consequent cost of, collecting this imperfect data, so the cost is something more than marginal.
- c) The subject scheme (Conspectus) used to describe the collections, is not used to describe items in the collection, and different collections use different schemes to describe their items, so once appropriate collections are identified, searches must be reformulated, often more than once, into other shemes.
- d) The data is not current. Snapshots are taken at yearly intervals at best. This may mislead the user in some instances.
- e) It is entirely possible that two collections are strong in a particular subject but that they largely or wholly duplicate each others' items so users may waste time in two collections when one would do.
- f) Strength measurements take no account of small collections that have unique items in them, users my be directed away from such unique items.
- g) The granularity level, whilst deeper than many used for measuring collection strength, is fixed and may be, at its deepest, still insufficiently deep for the users query, so that the user must 'guess at' the appropriate heading at the higher granularity level.

- h) Although the existence of a hierarchical approach compensates a little for this; the question of how the terminologies in users own minds map to the controlled terminologies used to describe collections and items in the system is unaddressed, and is likely to be a particular problem at granularity levels lower than those represented in the hierarchical subject menus offered.

In theory, locally built and maintained collection level description indices of the type described in annexe A.3 could be implemented at little cost for small and new e-resource collections, and could solve all of these problems except (h). The integration of the proposed HILT Phase II pilot terminologies server could solve this also as well as offering at least one additional advantage. Let's suppose the DNER, NOF and Digital Scotland set up 500 new e-resource repositories in the next 10 years and agree between them that each repository:

1) Will include the metadata listed below in item level records

- DDC number to a reasonably low level of granularity.
- The same subject term as everyone else uses for the number in question (through using HILT).
- Educational level based on MEG list.
- Unique object identifier.
- Data or format (e.g. electronic only).
- Charging policies (e.g. free, free to this or that collaborating group, educational cost, other cost, etc.).
- Service exclusion policy (eg permission to include in regional but not in national indices, etc.).
- Language(s) of content.

2) That each will built a collection level index containing these elements as a single string in an agreed order and make it available for cross searching via Z39.50².

Given appropriate facilities at the client end, the end result would be a dynamic landscaping facility that would solve all of the problems currently seen in CAIRNS. It would be

- Relatively objective³, based on a simple classification into subject categories like teaching, research, comprehensive, language, educational level, together with a simple numerical count.
- Relatively consistent in that a single approach would be applied across all services.
- Able to indicate duplication (using the unique identifier) and save their users' time in that respect.
- Able to show the existence of numerically weak collections containing unique items (unique identifier).
- Have up to the minute currency.
- Allow a single subject scheme and a single initial search to be used both at collection level and at item level in all collections.
- Allow collection strength description to the granularity levels used by users.
- Able to be created and maintained at marginal cost, despite being much better than the manual approach and much more current.

There would be metadata creation costs, of course, but in services where the use of this metadata is built in from the start, these costs would be offset against the need to provide the service's own users with this kind of information, the assurance that their users would

² Other solutions such as harvesting may be possible but might be less current.

³ and therefore more relevant and useful to the user.

have more current access to other services included in the approach where these were relevant, and the promotional value of having in-depth information on service content included in the collection level service. There would also be some system set up and maintenance costs, but these again could be regarded as offset by the advantages and would be marginal. Objectivity would not be absolute, of course. Assigning items to subject categories and educational level would be likely to have a subjective aspect, and there might well be a need to indicate the 'significance' of a collection or 'resource quality given a particular context' on something other than simple arithmetic in some instances. However, in theory at least, the objectivity level would be much higher, and it would be relatively consistent across all sites.

The service as described above would have limitations. It would not be able to address point (h) above. That is, it could not:

- Map the terminologies in users' heads to the DDC based collection strength measurement system

However, the proposed HILT Phase II pilot⁴ would solve this problem, together with various others relating to controlled terminologies. This pilot will allow:

1. Users to input terms then interact with the system to 'disambiguate' the term, perhaps against a DDC 'spine' (e.g. determine whether 'lotus' is lotus the car, the flower, the position, or whatever) – and also offer broader and narrower terms as alternative routes.
2. The mapping to DDC of 'user' terminology sets.
3. The monitoring of user terms over time to improve mappings over time.

If the user is directed towards use of HILT in the landscaping mechanism, it should be possible to utilise the above process to map user terms to DDC and hence to the distributed collection strength indices. Moreover, the use of HILT would also have at least one other advantage in relation to the landscaping mechanism. It would allow services who used HILT when assigning subject terms to items to assign their own terms as opposed to some centrally agreed set. As long as these terms were mapped in HILT to DDC and the user accessed the landscaping mechanism via HILT and DDC, this should not be a problem and would give sites more flexibility in meeting the needs of their own core users on their own site.

The same approach could work in the longer term for large legacy metadata sites, whether hardcopy or electronic as regards content, although it would require significant retroconversion programs for sites that did not or could not aim to be steady state sites that would ultimately drop older items from their collection. Such sites in particular (but also all legacy sites) might begin to deal with newly accessioned items in the new way as a first step towards this end. In the meantime, a simulated approach based on the SCONE idea of objectivised professional judgement could be tried. It would not be practical for this to reach the level of granularity that the automated system could, so that user queries at deep levels of granularity would have to be mapped upwards to a higher level for all manual sites. However, the use of DDC numbers and truncation makes this feasible and it should be possible to offer useful data by going to an agreed level of granularity generally in these manual assessments but also noting exceptions where there is greater strength or weakness in a given area.

⁴ See main SCONE report Appendix F for more information on this project.

A Two Stage Process

Of course, on its own the approach described would probably only be practical with a limited initial set of possible repositories such as those encompassed within the DNER or Digital Scotland or NOF. Given a bigger initial set, a two stage process would operate. In the first step, the total number of possible repositories would be reduced to a smaller, more relevant, set utilising data in the headings below, either singly or in combination, to define the initial sub-clump of repositories likely to be relevant.

Headings for Step 1

- Task (the JISC funded Copac/clumps project aims to look at the issue of tasks classification).
- Geographical Location.
- Sector (HE, FE, etc.).
- Domain (Museums, archives ...).
- Collaborative Groups (e.g. Glasgow Digital Library).
- Subject (would generate everything relevant up to and including 'general').
- Language.

The cross search of the proposed distributed collection strength indices would thus be limited to a relatively small group in most cases. Where even this group was too big for practical purposes (e.g. geographical location = the whole world or even the whole of the UK, a staged approach to widening out from practical levels of cross-searchable sites could be enforced).

Conclusion and Recommendation

What this sketch seems to show is that the idea of allowing users to 'scope ahead' in the DNER or Digital Scotland or NOF by using an extension of the current CAIRNS dynamic landscaping mechanism as described above is a feasible proposition worth investigating further. SCONE's recommendation is that consideration be given to funding a project focusing on new or small e-resource services that would:

1. Investigate the requirements in respect of the data content of local collection strength indices that would form the basis of the system.
2. Investigate the requirements of ensuring a standard approach to creating these indices and their content across large numbers of sites.
3. Investigate the requirements of simulating a similar approach for legacy sites.
4. Build and test a pilot system as described in this Annexe (A.4).
5. Examine scalability and architectural issues, focusing on criteria like task, geographical location, sector, domain, collaborative groups.