Do we really know what users want?

The DiSCmap and OCRIS projects

Gordon Dunsire

Presented at 13.seminar Arhivi, Knjižnice, Muzeji
25-27 Nov 2009, Rovinj, Croatia

Published by Gordon Dunsire

Edinburgh 2012

Do we really know what users want?
The DiSCmap and OCRIS projects

Presented at 13.seminar Arhivi, Knjižnice, Muzeji
25-27 Nov 2009, Rovinj, Croatia

Gordon Dunsire

Do we know what users want? Two recent projects conducted by the Centre for Digital Library Research at the University of Strathclyde, Glasgow, Scotland have produced evidence that information professionals are not taking into account user needs when providing access to digital resources. The two projects are DiSCmap and OCRIS; together, they reveal some differences between professional practice and user needs

DiSCmap

The Digitisation in Special Collections: Mapping, Assessment, Prioritisation (DiSCmap) project[1] was funded by the UK's Joint Information Systems Committee (JISC) from September 2008 to June 2009. The project was carried out in partnership with the Centre for Research in Library and Information Management (CERLIM).

The aims of DiSCmap were to:

- Identify priority UK Higher Education special collections for digitisation
- Assess user needs, across disciplines, for the digitisation of special collections
- Provide strategic recommendations to JISC

The special collections were identified by asking professional intermediaries such as archivists and librarians to nominate special collections from their local holdings. Users were also asked to nominate collections that they would prioritise for digitisation.

The responses from intermediaries were gathered from an online questionnaire which identified which local collections were a priority for digitisation along with factors relevant to digitisation such as the size and subject matter of the collection and the format of constituent items. The questionnaire also asked for reasons why digitisation was a priority, such as current and potential usage, the condition of constituent items, and current access arrangements. The responses to a similar survey carried out at the same time by RLUK (Research Libraries UK)[2] were incorporated into the DiSCmap results; overlap between the surveys was reduced by DiSCmap not notifying its questionnaire to RLUK members. The views of end-users were gathered from workshops and interviews with university students, researchers and teachers when they were asked which collections they wanted digitised. 945 collections were nominated in total.

An in-depth study of nominated collections located in Scotland was undertaken by matching them collection-level descriptions recorded in the Scottish Collections Network (SCONE)[3]. In general, only a collection title and location were given in the nomination. This was often insufficient to identify a match in SCONE, so additional information was sought from the Archives Hub, library websites, and

Google. A record for the collection was added to SCONE if no match could be found and sufficient information was available.

A collections "landscape" for nominated Scottish collections[4] was created in the Scotland's Information service, which uses the SCONE descriptions.

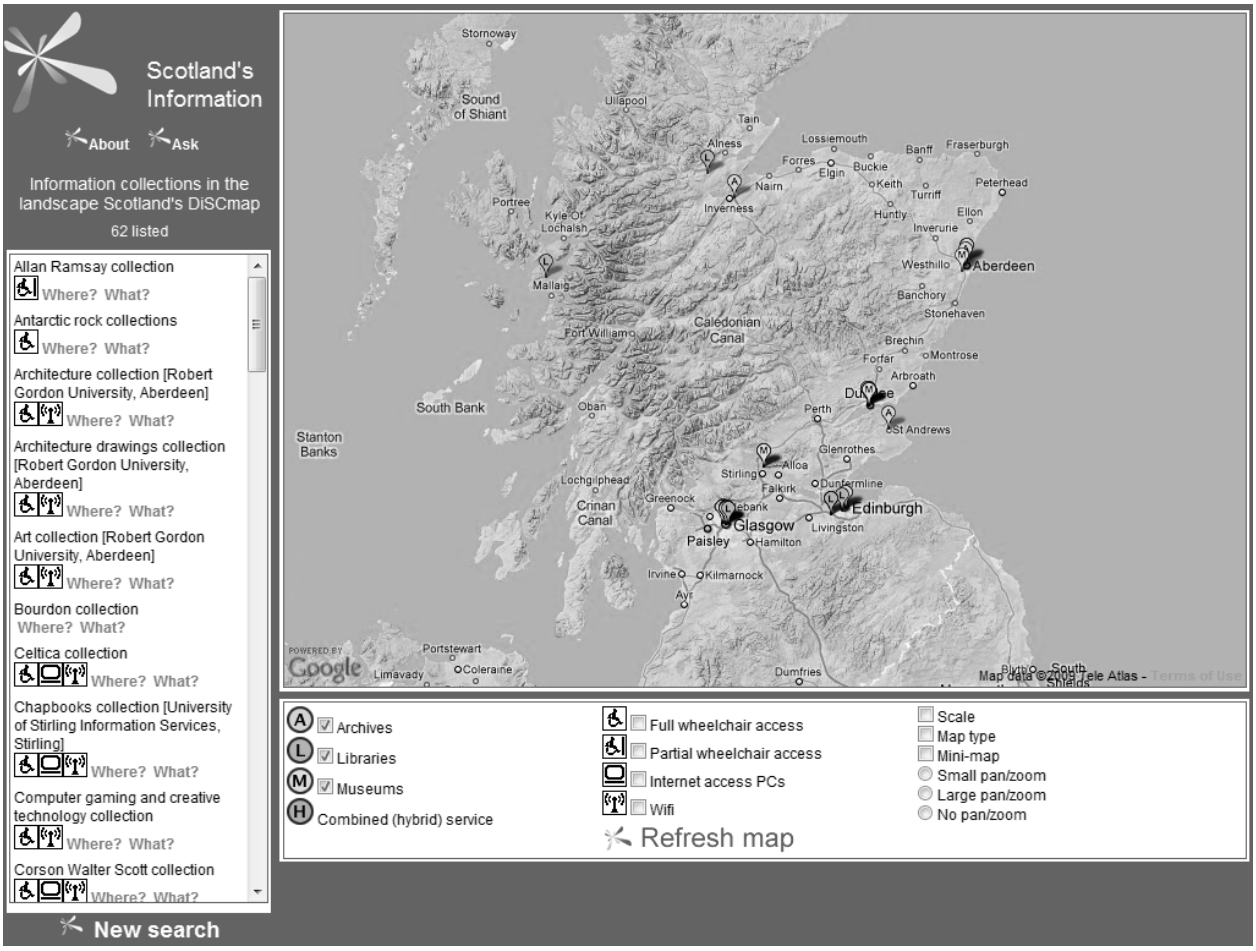

Figure 1: Screen-shot of the DiSCmap collections landscape in Scotland's Information.

The study of Scottish collections identified significant differences between the collection titles recorded in SCONE, which are based on online references and descriptions from library websites, and the titles given by intermediaries and end-users. There were also differences between the granularity of sub-collections in SCONE and those nominated in DiSCmap. Differences in titles and granularity were also noted across all nominated collections, and not just those located in Scotland.

User and professional perceptions of collections

Users often nominated large aggregations such as those at institutional level or super-collections of all special collections. An example of the former is "British Library"; an example of the latter is "Bristol University Special Collections". One user's reply to which collections should be digitised was "A vast array, across a wide range of subjects". These are not unreasonable responses, as users were not asked to take cost into consideration.

On the other hand, information professionals frequently nominated sub-collections of special collections, often comprising aggregations of items in a specific format within a named collection. Examples include "Maps from the Gough Collection", "Edward Clark Collection: glass slides", and "Gallacher Memorial library. Pamphlet collection".

There were also differences in the location of collections. Users were happy to nominate collections at super-institutional level, or located outside the UK: "Libraries holding medieval manuscripts", "Akademie der Kunste Berlin", and "Archives du Nord, Lille" are examples. But professional responses to the DiSCmap online questionnaire rarely mentioned non-local collections to complement a nomination. They were not specifically asked to do so, although several nominated collections were clearly described as being part of a larger collection distributed over different institutions in several locations. The RLUK survey did ask for complementary collections to be mentioned, but the responses within that survey were only partly reciprocal – Library A might mention Library B's collection as a complement to its own, but Library B did not necessarily mention Library A's collection.

The question "What is a special collection?" arose at an early stage of DiSCmap, during testing of the online questionnaire for professional intermediaries. The project adopted a pragmatic approach; if a librarian or archivist called something a special collection, then it was treated as such. This was taken after various published definitions were considered – defining factors included rarity, physical form, subject, depth of coverage, and intended audience amongst others. The question is closely allied to "What is a collection?", and the same pragmatic answer is used, for example, by SCONE and the Dublin Core Collections Application Profile[5]. Like SCONE and Dublin Core, the DiSCmap definition of "collection" is based on Michael Heaney's analytical model of collections[6].

Users seem to have a broader view of collections. They tended to nominate big collections, and ownership and location appear to be of little concern. This is hardly surprising, as users presumably are interested primarily in the content of a collection. Location ceases to be factor affecting access when the content is digitised. Ownership may not be perceived as an issue if the user is allowed free access to the original collection in situ. Users may also disregard licensing as important if they assume that digitisation usually leads to more, rather than less, open access. And where subscription or licensing fees are charged, there may be an assumption that they will be paid for by the user's institution or library.

Professionals tend to have a narrower view of collections. This may be because they are more aware of professional reasons for prioritising materials for digitisation, including their rarity, fragility, handling requirements, and copyright status. Format-specific issues seem to be of much greater concern to professionals than users. Many of the professionally-nominated format-defined sub-collections involve formats such as pamphlets, manuscripts and photographs which are difficult to curate. Professionals are also more likely to be aware of, and concerned about, the cost of digitisation.

These differences in perception of collection granularity will have little impact when the collection itself is scoped on a specific format. Special collections containing only one material format are common: photograph collections and manuscript collections are examples. However, if the collecting scope results in mixed formats, as for example with a subject-focussed collection, then

user perceptions of what should be digitised may well differ significantly from the opinions of information professionals.

Lack of professional agreement on the definition of "special collection" and "collection" spills over into the question "What is the title of the collection?". The formal title of a collection is important for searching, listing and identification. This functionality is impaired by the prevalence of generic titles in use, often based on material format: "Pamphlet Collections" is the title of four separate nominations for DiSCmap, and "Incunabula" or "Incunables" of five. While this may be excused by the lack of a national framework for recording collections in the UK, which could provide guidelines for constructing unique titles, the lack of consistency evident from the in-depth study of Scottish collections is mysterious. The study found many instances of variation in titles for the same collection used by its holding library or archive. For example, a DiSCmap nomination for the "Glasgow School of Art photographic collection" is a variation of the title already recorded in SCONE as "Glasgow School of Art photographs collection". The SCONE title would have been derived from the best available information from the institution's website when the collection-level description was created. However, further investigation found two current variants on the website, "GSAA P: Glasgow School of Art photographs" and "GSA Archive Photographs", neither of which exactly match the SCONE title. All four variations refer to the same collection, but there is no set of keywords that can retrieve them all, even with full truncation. Which of these is the "correct" title? Which of these should be used to reference the collection?

Variations in title were found to affect 10 percent of nominated collections already in SCONE. Perhaps users are wise to nominate collections that have most or all of the materials in a library or archive: they are ensuring that the "right" stuff is digitised.

OCRIS

The Online Catalogue and Repository Interoperability Study (OCRIS)[7] was also funded by JISC as a three-month project ending in September 2009. The aims of the project were to:

- Survey any overlap in scope and content between Online Public Access Catalogues (OPACs) and Institutional Repositories (IRs).
- Examine interoperability between them.
- List services to managers, teachers and learners.
- Identify actual and potential links to administration systems.
- Make recommendations for improvement.

The survey took the form of an online questionnaire, desk research of organisational websites, and two in-depth case studies at Cambridge University and Glasgow University.

For the questionnaire, OCRIS identified 21 types of item that might be held in both the institutional repository and library of an organisation. These types were based on the Eprints type vocabulary encoding scheme[8] developed for the Scholarly works application profile. Analysis of the responses showed an 80 percent overlap between item types held in IRs and OPACs.

Desk research uncovered frequent duplication of metadata at record and item level. For example, electronic versions of theses are often stored and described in the institutional repository, while print versions are held by the library. Separate metadata records result in considerable duplication, as only the data at the manifestation and item levels of the model of Functional requirements for bibliographic records[9] will be different, while that at the work and expression levels it will be the same. Institutional repositories usually record journal articles, including pre-prints and refereed papers, at article level, while library catalogues typically provide metadata at the level of journal title or issue. But libraries frequently provide access to abstract and indexing services which also offer article-level metadata, and thus duplicate much of the data for this type of material in the institutional repository.

The survey identified very little interoperability between institutional repositories and online library catalogues. Only 2 percent of responses indicated any kind of current interoperability, while 11 percent claimed that interoperability was "pending" – that is, included in formal plans and strategies. Over 85 percent of organisations have no actual or planned interoperability.

OCRIS identified two main ways that interoperability might be improved. Resource discovery platforms like Aquabrowser and Primo can provide interoperability within an organisation. Physical or distributed union catalogues, for example based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Z39.50 protocol respectively, can provide interoperability between organisations.

Both approaches are highly dependent on the use of authority control for headings, so that similar resources can be brought together and disparate resources kept separate during information retrieval. If there is no common authority file shared between the institutional repository and library catalogue, then a mapping between individual files, if they exist, is required.

Authority control in library catalogues tends to be highly developed with good support for format and content standards. But there is little agreement between libraries on using a common authority file. Even when a common approach occurs passively, for example through the employment of the same bibliographic records service, most libraries tend to manage authority files locally, and will include local name and subject headings which are not part of, or subsequently added to, the global file. Many institutional repositories do not use authority files: one third of those responding to the questionnaire do not use a name authority file; one third of those responding do not use a subject heading or classification scheme.

This suggests that institutional repositories are not focussed on information retrieval functions, or if they are, then their size is small enough to allow effective retrieval without authority control. Indeed, the primary purpose of many repositories is to support the funding of research by providing evidence of volume of research output, lists of high-ranking papers, and other metrics to inform, for example, the Research Excellence Framework (REF) in the UK.

Yet the potential users of library catalogue and institutional repository metadata include internal and external researchers, funding agencies, governments, teachers, learners, and the general public. The lack of interoperability between these metadata means that such users have to search in more than one place: multiple headings if the catalogue and repository can be searched via a union

catalogue or resource discovery platform, or separate search interfaces if they can not. In both cases, search results may need de-duplication.

The irony of this situation is that, generally, libraries have spent many years developing the use of standards to improve interoperability between library catalogues. OCRIS found that the library is often involved in creating or augmenting the metadata in its institution's repository, but this rarely extends to full authority control. One explanation may be the lack of simple, easy-to-use procedures and workflows. A surprising number of repositories use captions from the Library of Congress Classification as subject headings (with a certain amount of confusion about these being the same as Library of Congress Subject Headings), but this is because the classification scheme is bundled with the major repository software packages with a user-friendly hierarchical browse interface.

The intervention of the library community does not always lead to immediate improvements. OCLC went into partnership with the OAIster service[10] in 2009. OAIster harvests metadata records from institutional repositories around the world into a union file via OAI-PMH. OCLC maintains WorldCat, the world's largest union catalogue of library records. The OAIster records were added to WorldCat.org[11], the web interface to the union catalogue, with immediate negative effects on retrieval by name; while name authority control is well established in the WorldCat database, it is non-existent in the OAIster file.

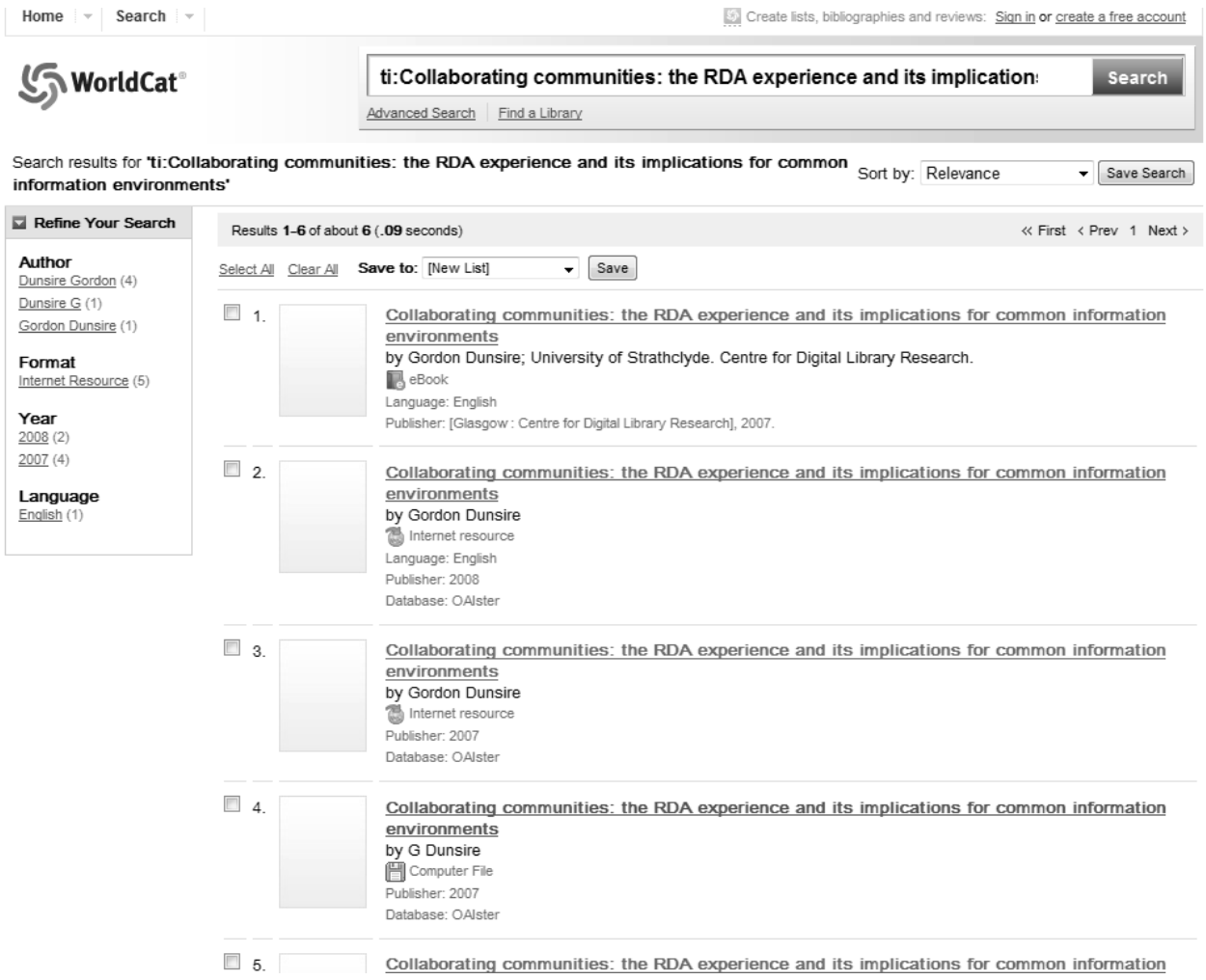

Figure 2: Partial screen-shot of worldcat.org.

The example in figure 2 illustrates the problems of duplicate records and lack of name authority control. The publication in question is the original English version of a paper published in the proceedings of AKM11. The English version was self-published and catalogued directly into WorldCat (in MARC21 format). A copy was deposited in E-LIS[12], a shared (inter-institutional) repository for papers in library and information science, with a separate metadata record in Dublin Core format. Another copy was subsequently deposited in Strathprints[13], the institutional repository for the University of Strathclyde, again with metadata in Dublin Core. Presumably the metadata records for either or both of E-LIS and Strathprints were subsequently harvested into another union file which was then re-harvested by OAIster; the result is 6 duplicate records for the same resource in WorldCat.org.

In the top left corner of the screen-shot there is a "related works" search titled "Refine your search". All three entries in the author section are for the same person, but only one can be selected at a time. The user wishing to see all of the metadata associated with this person will have to undertake at least three separate searches using this facility.

OCLC is attempting to alleviate these problems, but the current situation is a good illustration of the problems encountered when metadata is not rich or standard enough to support interoperability on a wide scale.

Conclusion

These findings from DiSCmap and OCRIS suggest that users are seeing a bigger picture than information professionals. Users seek information beyond the local collection and beyond the local institution. Yet user needs for easy-to-use discovery and access services are not being met by professionals; indeed, the situation may be deteriorating rather than improving.

Focus on deposit and preservation rather than retrieval and short-term institutional requirements may be preventing improvement.

In 2005, an RLUK study recommended further investigation of user needs[14]. Some significant work has been carried out since, notably by the Research Information Network[15], but as yet there is little evidence of better services as a result.

Is this what our users want?

References

[1] DiSCmap: Digitisation in special collections: mapping, assessment, prioritisation. Available at: http://discmap.cdlr.strath.ac.uk/
[2] RLUK: Research Libraries UK. Available at: http://www.rluk.ac.uk/
[3] Scottish Collections Network. Available at: http://scone.strath.ac.uk/Service/Index.cfm
[4] Scotland's DiSCmap landscape. Available at: http://www.scotlandsinformation.com/SIShow.cfm?TI=17&ST=1&MC=sca,typ,ove,pzl
[5] Dublin Core collections application profile. 9 March 2007. Available at: http://dublincore.org/groups/collections/collection-application-profile/
[6] Heaney, M. An Analytical model of collections and their catalogues. 3rd issue, revised. 2000. Available at: http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf

[7] OCRIS: online catalogue and repository interoperability study. Available at: http://cdlr.strath.ac.uk/ocris/
[8] Eprints type vocabulary encoding scheme. Last modified 14 May 2008. Available at: http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Type_Vocabulary_Encoding_Scheme
[9] IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records: final report. As amended and corrected through February 2009. Available at: http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf
[10] OAISter database. Available at: http://www.oclc.org/oaister/
[11] WorldCat. Available at: http://www.worldcat.org
[12] E-LIS: E-prints in library and information science. Available at: http://eprints.rclis.org/
[13] Strathprints: University of Strathclyde institutional repository. Available at: http://strathprints.ac.uk/
[14] Digitisation in the UK: the case for a UK framework: a report based on the Loughborough University study
on Digitised Content in the UK Research Libraries and Archives Sector commissioned by JISC and the Consortium of Research Libraries (CURL). Version 1.1. November 2005. Available at: http://www.rluk.ac.uk/files/Digitisation_in_the_UK.pdf
[15] Research Information Network: project reports (A-Z). Available at: http://www.rin.ac.uk/resources/rin-publications/project-reports