Copac Collection Management Tool Project

Leeds case study 1

Large scale use of the Copac Tool to profile collections

Background

Around 4 years ago Leeds University Library decided to begin a programme of profiling its collections in order to assist with collection management decisions. By identifying those subject collections - "heritage collections" - which might be considered significant within a national or international context, these could then be prioritised in terms of allocation of scarce resources such as space, preservation and conservation, ongoing collection development. Such collections would be seen as a long-term asset capable of enhancing the reputation of the Library and of the University and of attracting researchers to Leeds. In contrast, "self-renewing collections" would be maintained for as long as they aligned with the University's needs in terms of learning, teaching and research, but could be considered expendable should they cease to do so because of a change of focus in the University's activities[1].

Heritage collections were therefore defined as those which meet one or more of the following criteria:

- A rare or unique collection i.e. of national or international significance
- The collection is closely associated with a significant non-print collection in the Library e.g. manuscript/archival/digital resources so that together they represent a unique and valuable resource.
- The collection deliberately brings together items normally dispersed around other UK HE libraries, thus increasing the overall significance and strength.
- The collection pertains in some way to high profile work associated with the University, or to Leeds or Yorkshire more broadly

The first and third of these criteria relate to the broader picture of national holdings in the pertinent subject area, and can therefore be informed by holdings information from Copac.

Initially, the only possible approach was to collect expert knowledge and opinion from library staff – both current and retired, as well as academic staff. However we quickly began to question the subjective nature of this approach, especially once the Copac Tools became available and were seen to offer a sounder and evidence-based approach for this programme. Even so, the initial gathering of statistics itself needs to be based on a knowledge of our collections and will need to be followed by an expert assessment capable of bringing contextual knowledge to the interpretation of the evidence in order to assign each subject collection to the appropriate category.

[1] Full details are in the Library's Collections Strategy, available at: http://library.leeds.ac.uk/downloads/file/212/collection_strategy

Using the Copac Tool : the method adopted

The decision was taken to base our analysis on the breakdown of the collections offered by our classification scheme. In other words, a collection would be defined as the books classed within a particular class range. This was made easier by the nature of the Leeds classification scheme which uses the names of subject disciplines such as “French” or “Electrical Engineering” at the top level of the classification hierarchy.

However, the option does exist, where necessary, to combine two or more specific classmarks to form a single set of of titles for submission to the Copac Tool where this provides an appropriate grouping. For example, Leeds has a strong collection about the Orthodox Church, split between two main areas in the scheme, one concerning the organisation, doctrine and liturgy of the Orthodox Church, the other concerning its history.

It is also possible to analyse at different levels of specificity. So each subject within our scheme has been treated as a whole, but sub-disciplines have also been profiled, based on the subdivisions within the classification scheme. This dual approach gives additional evidence about the precise areas of strength and therefore assists with the subsequent process of profiling.

Once the scope of a subject for analysis has been defined, a query is run in our Millennium Library Management System which identifies the relevant bibliographic records. The ISBNs are exported from those records and submitted to the Copac Tool via the batch search facility. Matching records are identified and results are then always de-duplicated by ISBN. This process provides a way of allowing for the fact that there are many titles within the present Copac database for which duplicate records exist: provided all such records contain the ISBN submitted for the title (which would usually be the case), then the results will treat all the holdings as if associated with a single record.

The resulting graphs are analysed and the following data elements recorded:

- % of records within the collection held by 3 libraries or less (including Leeds)
- % of records within the collection held by 4 libraries or less (including Leeds)
- Number of libraries holding 2/3 of the titles within the collection
- Number of libraries holding 1/2 of the titles within the collection
- Number of libraries holding 1/3 of the titles within the collection

It was hoped that by recording all these various data elements, patterns would emerge which would allow us not only to identify our “heritage collections” but would also allow us to define a set of generic parameters normally characterising such collections in order to permit great confidence in interpreting the results as the project progresses.

The one real limitation up to present has been that the searching needs to be carried out by ISBN rather than by record number, otherwise only the records with Leeds holdings attached would be retrieved. Given the high number of duplicate records within Copac, the results would completely lack validity as they would fail to be based on all copies of that title recorded in Copac. This limitation has constrained Leeds to

only analyse data for subjects where the majority of the titles – or at least those which are most significant – have records with ISBNs. In practice, this has largely restricted us to subject areas in medicine, science and engineering, although some interesting data have also been obtained for selected subjects in the arts and social sciences (while recognising that such results only relate to recent acquisitions rather than the whole collection).

Results

It should be noted that:

- Data regarding file size is missing for the earliest sets submitted.
- If a collection contained more than 7,000 records, then the file was reduced to an appropriate size by sampling in order to allow faster processing and display within the Copac Tool. The sampling was achieved by taking every n^{th} record from the file in shelf order so that the sample accurately reflected the spread of subject matter across the collection.
- Any sub-discipline with fewer than 100 titles has been omitted from the following table as being too small to allow conclusions to be drawn. In any case, any such collection is unlikely to be heritage because that categorisation implies a breadth which would certainly be lacking from a collection of that size.
- We realised that we should limit our review file to capture material published up to the end of 2009. Otherwise material may seem rarer than it actually is because of delays in some libraries in cataloguing material or in uploading their records to Copac.
- In a few cases, specific sub-disciplines (e.g. Modern French Literature) were analysed in isolation in order to explore particular relationships and comparisons.

Subject/classmark	File size	Libraries holding 1/3	Libraries holding 1/2	Libraries holding 2/3	% in ≤3 libraries	% in ≤4 libraries
Chemistry – All		19	8	5	8%	11%
Chemistry – Analysis	199	11	7	4	10%	14%
Chemistry – Theoretical	131	22	13	6	4%	7%
Chemistry – Spectroscopy	169	16	11	6	2%	4%
Chemistry – Properties of matter	470	14	11	4	7%	11%
Chemistry – Chemical kinetics	158	18	10	4	4%	6%
Chemistry – Inorganic	295	24	11	6	4%	6%
Chemistry – Organic	949	22	10	6	9%	12%
Colour Chemistry - All		8	6	2	25%	32%
Communications Studies – All	3582	16	8	3	11%	15%
Communications Studies – Media	919	15	7	3	13%	18%
Communications Studies – Journalism	262	9	7	2	18%	24%
Communications Studies – Social aspects	1414	20	8	3	9%	12%
Communications Studies – Political aspects	881	14	8	4	10%	15%
Computer Studies – All						
Computer Studies – Systems Organisation	749	15	10	5	4%	7%
Computer Studies – Software	2202	14	7	4	9%	11%
Computer Studies – Data	167	14	11	9	2%	5%
Computer Studies – Theory of computation	767					
English – Modern	4978	17	8	7	4%	6%
French – Modern	3452	8	4	3	20%	27%
German – Modern	3347	8	6	1	25%	30%
Health – Anatomy	422		8	6	7%	11%
Health – Physiology	274	20	12	6	7%	8%

Health - Biochemistry	414	21	12	4	8%	10%
Health – Pharmacology	885	31	29	29	9%	14%
Health – Microbiology & Immunology	474	23	14	5	6%	7%
Health – Clinical pathology	104	16	8	3	11%	15%
Health – Pathology	791	15	5	4	12%	19%
Health – Medical profession	2889	27	14	6	6%	8%
Health – Public health	4278	24	13	4	7%	10%
Health – Practice of Medicine	1210	16	10	6	5%	8%
Health – Communicable diseases	769	19	7	3	10%	13%
Health – Nutritional & metabolic disorders, etc.	199	18	8	6	9%	13%
Health – Musculoskeletal system	903	14	9	5	12%	15%
Health – Respiratory system	358	16	10	5	10%	13%
Health – Cardiovascular system	809	13	8	4	11%	15%
Health – Haemic & lymphatic system	298	14	8	3	7%	10%
Health – Gastrointestinal system	431	14	6	3	7%	12%
Health – Urogenital system	352	14	8	4	11%	16%
Health – Endocrine system	301	14	9	4	7%	10%
Health – Nervous system	1453	19	9	5	6%	10%
Health – Psychiatry	3714	17	10	5	9%	12%
Health – Psychology	848	25	12	6	7%	10%
Health - Radiology	771	11	8	3	14%	20%
Health – Surgery		18	10	6	4%	7%
Health – Gynaecology		11	7	3	13%	20%
Health – Obstetrics		16	12	6	7%	9%
Health – Dermatology		16	9	5	11%	12%
Health – Paediatrics		19	11	5	7%	10%
Health – Geriatrics		20	13	5	7%	9%
Health – Dentistry		17	13	4	7%	10%
Health – Otorhinolarynology		13	8	3	12%	19%
Health – Opthalmology		17	11	6	5%	10%
Health – Hospitals		17	9	4	9%	12%
Health – Nursing		15	14	6	10%	14%
Health – Bibliography & reference		24	13	6	8%	10%
Icelandic – All	702	7	5	4	25%	40%

Management - all		13	8	4	10%	14%
Modern History – Russian history		9	6	5	12%	20%
Orthodox Church	523	5	5	4	21%	31%
Physics – All	2861	20	9	4	6%	9%
Physics – History of physics & natural philosophy	182	24	13	8	5%	9%
Physics – Physical properties of matter	270	16	6	4	9%	12%
Physics – Solid state	538	19	8	4	7%	11%
Physics – Quantum	527	20	8	5	6%	8%
Physics – Acoustics	1070	19	9	4	6%	9%
Physics – Optics	342	20	9	5	6%	9%
Physics – Heat	170	25	13	8	3%	5%
Physics – Electricity & Magnetism	373	18	10	4	7%	13%
Physics – Geophysics, meteorology, atmospheric	104	15	9	5	11%	14%
Social Policy – All		25	19	6	5%	8%
Social Policy- Disability		19	9	4	11%	17%
Transport All		11	6	3	16%	20%
Transport – Highways	182	11	7	3	18%	21%
Transport – Traffic management	254	9	4	2	21%	27%
Transport – Transport planning	610	15	8	5	11%	16%
Transport – Ground transport	244	9	7	5	13%	16%
Transport – Air transport	268	7	5	2	34%	38%
Transport – Transport economics	433	15	4	3	9%	13%
Transport – Freight	110	10	7	5	14%	16%

Analysis and discussion of the results

When the results of this exercise so far are compared with what we know about our collections, their history and recent collecting policy, we believe that the following collections stand out as potentially heritage:

Subject/classmark	Libraries holding 1/3	Libraries holding 1/2	Libraries holding 2/3	% in ≤3 libraries	% in ≤4 libraries
Colour Chemistry – All	8	6	2	25%	32%
French – Modern	8	4	3	20%	27%
German – Modern	8	6	1	25%	30%
Icelandic – All	7	5	4	25%	40%
Communication Stud. – Journalism	9	7	2	18%	24%
Orthodox Church	5	5	4	21%	31%
Transport – All	11	6	3	16%	20%

Previous work in the first phase of the project had suggested that our collections in Colour Chemistry are exceptional[2]. It is gratifying that, with a much wider range of collections now analysed, this still seems to be the case, although the omission of data for the older titles does mean that a significant dimension of that collection has still to be explored.

The results for French, German and Icelandic are interesting. We expected Icelandic to emerge very strongly as we, along with UCL, are noted for our holdings in this area. However the results showed a significant overlap also with the British Library and Cambridge, so the question would be whether that result would be replicated if we were able to include older material. More generally, for subjects where much of the material is literary, it is still not clear how strongly the results are influenced by the patterns of publishing and purchasing, which differ significantly from those for non-literary subjects. The low percentage of rare material in our modern English collection does suggest that the figures for French and German may reflect real strengths or, at the very least, that the Leeds collections are not widely replicated around the country. Once again, inclusion of older material, particularly when considered in the context of the results for the subject area as a whole may well alter our perceptions regarding all 3 subjects.

The relative rarity of the material on journalism (albeit the figures are based on a very small number of titles) did come as a surprise. It may be that similar collections are often located in institutions which do not contribute to Copac, and the results are therefore somewhat misleading regarding the national picture. It is also interesting to view the results across the whole subject area of Communication Studies and to compare these with more traditional subjects such as Chemistry or Physics.

The Orthodox Church "collection" was interesting in that although it represents a logical "discipline", it brings together titles from two separate areas of the

[2] See the discussion on pages 18-23 of the Phase 1 final report at http://copac.ac.uk/innovations/collections-management/wp-content/uploads/2012/01/Copac-Collection-Management-Tools-Project-Final-Report.pdf

classification scheme and represents a very distinctive collection. It emphasises the importance of considering relationships between subjects, an area on which Leeds intends to undertake further investigation. We had expected Russian history to also show as heritage; however the evidence from the Copac Tool suggests that it is marginal. This may reflect the inability to include older materials, the fact that Russia was late in adopting ISBNs, and also reduced expenditure on this subject Sarea in recent years, and so we should wait until we are able to process material without ISBNs before forming a definitive view.

Our collections in the area of Transport Studies were already known to be strong. The results confirm this, though perhaps less clearly than we had originally anticipated. The collections also include a quantity of grey literature, much without ISBNs, and inclusion of that material might give a more clear-cut result.

We were surprised that Radiology and Gynecology were also both borderline – we were not aware of any particular strengths in these subjects. Detailed investigation of the unique titles in these two areas would be useful in clarifying the underlying reasons.

In all of this detail, what is reassuring is that we can clearly see figures suggesting that the bulk of our collections in science and medicine are not exceptional, even though they may be very strong working collections. Equally, there is manifestly a distinct group of subjects from very different fields where we have always felt that the collections were distinctive and where this is borne out by the results.

It is therefore possible to suggest tentatively that a collection might potentially be of heritage status if it matches one of the following criteria:

- 15% or more of the titles are in 3 libraries or less
- 21% or more of the titles are in 4 libraries or less
- 2 or fewer libraries hold 2/3 of the titles

Further results are needed before this can be considered anything other than a working hypothesis – but this in itself is a tremendous step forward for us.

Current limitations

Some caution is required in interpreting results because of limitations intrinsic to the use of the Copac database:

- Additional copies of relevant titles may be held by other libraries that are not Copac contributors
- Equally, the results cannot take account of material held by Copac libraries but which is currently uncatalogued or held in specialist catalogues and not uploaded (e.g. the East Asian holdings of some RLUK libraries)
- Differences in record quality or cataloguing practice can result in a failure to match identical records (e.g. some substandard records do not contain an ISBN).

More fundamentally, any search not based on ISBN will currently miss all duplicate records for the same title. The vast majority of the Leeds collections which we think

are probably ‘heritage’ are those which contain older material that we have not been able to test so far. These are mainly in the Arts, but there are also some science subjects, and a large quantity of foreign language material, some of it from countries late to adopt ISBNs. For these, there is currently no value in attempting to use the Copac Tool as the results will not give an accurate picture of national holdings – although this should change when Copac migrates to its new platform.

It would therefore be advantageous for the Copac Tool to be developed to handle non-ISBN material appropriately. This could be achieved via “search expansion”, with the following sequence of steps comprising a batch submission:

1. Submit a list of record numbers to the Copac Tool.
2. The bibliographic data is analysed for each record selected and author/title/date searches submitted for each record (perhaps similar to the former 4,4,date key).
3. De-duplication of the resulting results set based on the options already existing within the Copac Tool.

Conclusions

This case study demonstrated the value of the Copac Tools in supporting the collection categorisation work currently being undertaken at Leeds as a key element of the Library’s broader collections strategy. After running more than 80 collections and sub-collections through the Tools, a degree of consistency is apparent in the results which gives us confidence both in the process involved and in its ability to highlight those collections which are worthy of further examination. Put simply, it has enabled the process of categorisation to change from a very subjective and potentially controversial exercise to one underpinned by solid statistical data, even though the ultimate decisions still call for professional expertise and an intimate knowledge of the collections.

As the Collections Strategy document already cited makes clear, the categorisation of collections will permit the Library to identify those collections which are unique and distinctive, make best use of its resources for acquisition, subscription, storage, preservation and digitisation, and place its curation of collections on a secure and cost-effective footing in the long-term.

At present, the difficulties associated with searching by record number hinder analysis of the Library’s older materials. This restriction prevented us from gathering the information required for a fuller exploration of precisely what statistical profile would be suggestive of a heritage collection and whether there is any variation by subject area. It is important also to remember that any results are based on the data held within Copac; staff knowledge of additional collections either not yet catalogued or held in libraries that do not contribute to Copac may well be relevant in forming a final judgement. However the value of the Copac Tools for an exercise of this kind is amply demonstrated by the work undertaken so far.

Michael Emly, Jonathan Horne and Maureen Pinder, July 2012