Copac Collection Management Tools Project Final Report

31th July 2011

Table of Contents

1. Introduction

This report is the second of two companion project reports. The first, the "Interim Report – 31st May 2011"[1], details the progress of the Copac Collections Management Tools Project through its initial and second phases. These principally focused upon the development of a prototype tool to support collection managers in the decisions they need to make around monograph retention and disposal and collection development. It included details of the observations produced through extensive testing, together with descriptions of a series of use case scenarios developed by the project team looking at workflow and accruing benefits associated with the application of the CCM Tool.

This final report provides a description of the work undertaken in Phase 3 of the project. It builds upon the earlier work of the project, describes further technical development and subsequent testing of the prototype Copac Collection Management Tool. The development of the four use case scenarios produced in the Interim Report have been added to with a further two detailed in this report within section 6. Section 7 takes use case two, "Identifying Collection Strengths" as the basis for a detailed case study, providing a practical demonstration of the current potential for the CCM tool and also confirming areas requiring further development.

A "Description and Overview" is articulated in the Interim Report which is not repeat here, however it is important to note that the opportunity to pursue this JISC supported seven month project has resulted in a reinforcement of the commitment amongst the project partners, realising how important the development of the CCM tool is. If developed further, this tool will make a real difference to helping universities make informed decisions particularly at a time of financial constraint. The project has clearly demonstrated the value of exploiting community data (Copac), enabling more effective and efficient approaches to monograph collections management, the benefits of which the project has begun to articulate through the use case scenarios.

The project has been funded as part of the broader JISC & RLUK RDTF/Discovery initiative. The project specifically addresses one of the key aims of the RDTF Vision: Support collection management processes and reduce duplication of effort. (See: http://discovery.ac.uk/files/pdf/jisc-rluk-vision-final-june2010.pdf)

2. Project Aims and the Phase 3 Goals

The CCM Tools Project started in January 2011 with the principle goal of providing a proto-type of a web based collection management support service for academic libraries using Copac data, incorporating a variety of tools to support library staff in collection development. This has been achieved. An Interim Report was produced in May 2011 that described in detail the progress to that point including details the iterative development and testing undertaken; the collective thinking around benefits and use case examples of the CCM Tool; and a look forward to next steps within and beyond the project timeline.

As described in the Interim Report:

---

[1] http://www.rluk.ac.uk/files/CopacCMInterimReportfinal.pdf

"The overarching aim of the project is to develop and test a service that will enable improved decision making regarding the retention, disposal, and redistribution of materials. The service will provide evidence of the wider availability of individual materials and/or collections when discussing the disposal of materials with academic staff within an institution. This project will help RLUK, JISC, and the WRC achieve the longer term aim of developing the technical framework required to support a more proactive and cohesive approach to collection management at a national level. In addition, by building the service on top of Copac data, the project contributes significantly to furthering the work of the JISC & RLUK Resource Discovery Taskforce, which aims to explore how data can be opened up and made to 'work harder.'"

The two months since May has seen the completion of the project's Phase 3 activity. The project set its goals high, not only to complete the tasks assigned but as working as a cohesive team to explore the possibilities of the further development of the CCM Tool and the benefits that may accrue from its application. The Use Case scenarios and case study are an important part of this thinking and section 6 and 7 details some further work done in this area. The testing of the CCM Tool looked beyond the current project to enhancements that would be beneficial if the tool was to be developed further. Section 9 explains this further.

3. Methodology

The Copac Collection Management Tools Project was made up of three phases.

Phase 1 – The development of a trial collection management interface for Copac
Phase 2 – Iterative testing and feedback of the tool by the White Rose Consortium partners

These two phases were the subject of the **"Interim Report – 31st May"**, and were successfully completed to schedule. They generated findings and observations that were taken forward and used to inform the technical develops in Phase 3 and the testing regimes.

A number of technical developments that were identified as desirable were beyond the scope and time limits of the Phase 3 activity and these have been held over to inform potential further phases of activity for the CCM Tools Project. A summary can be found in section 9 below with other detail contained within the Interim Report, section 6.

The Phase 3 work that ran from June to the end of July 2011 can be broken down into the following areas:

Phase 3	Further options resulting from Phase 2 findings detailed in the interim report.
a. Revision of Interface	Based on feedback make agreed changes to interface.
b. Batch Search	Identified as a desirable option providing a facility to upload a file of standard numbers for batch searching, with RSS update on result availability. This includes the option to do a range of different types of search based on the content of the identified records, offering a range of match confidence. Deliverables: • Facility to upload a file of local record numbers

	(for RLUK members) or a file of record identifiers such as ISBN/ISSN. Initial limit on the number of records in a set; to be determined. - Provide a name for the batch. - Select from a range of search types eg. ISBN. - Provide batch search RSS feed update. - Provide a result set including details of the number of locations and records for a document. - Result visualisation options as for direct search.
c. Subject Search	Provide a subject search option allowing users to explore the holdings of libraries across the country in particular subject areas. Deliverables: - Facility to search for a subject. - Provide a result set including details of the number of locations and records for a document. - Result visualisation options as for other searches.
d. Search Limit	Provide a range of search limit options and assess the potential use of a document retention level indicator. In the first instance we would offer the option to limit a search (direct search or batch load) by: - option to limit a search (direct search or batch load) by the following: Individual library; Region. - Agree a scale of retention levels with definitions of those levels and discuss the potential for limiting results by retention status.
e. User assessment and user testing reports	The testing of implemented options carried out throughout Phase 3 and at the end to gain overall feedback on the test interface. Views gathered on how the work might be built upon in the future and what priorities for further development should be. Phase 3 user testing reports will be produced.
f. Final Report	The final project report including issues and recommendations.

The project Partners have established an effective cooperation with an efficient division of labour. Clear responsibilities and task assignment have ensured that the project schedule was easily maintained over the duration of the seven month project. Responsibilities have been delineated as follows:

- Mimas - lead partner, responsibility for the overall project and in particular for the development of the CCM Tool.
- The University of Leeds - lead contact for the White Rose Consortium partners (Leeds, Sheffield and York universities) whose principal role is the testing and feedback of the CCM Tool, looking at how it integrates into workflows and developing use cases.
- RLUK - leading on external communications and outreach.

An agreed Partner Agreement is in place.

3.1.   Dissemination and Outreach

Dissemination and outreach has been a consideration throughout the duration of the project. During the final two months an article was accepted by Inspire to appear in their Newsletter. This is emailed directly to a wide range of people via the following channels: Lis-Inspire; the SCONUL email list and the Society of Chief Librarians list. It is also published via the Inspire website: http://www.liem.org.uk/Inspire. A project update entitled "Getting Excited about Collection Management" has also appeared via the Copac Development blog which has been relayed via Tweets on RLUK. The Interim Report was made available on the RLUK website.

4. Technical Development

The Phase 3 technical development work had a slightly different timeframe from the rest of the Phase 3 activity. In order for the testing to be successfully completed work began on the development of the batch functionality of the interface in April. On the 27th May Copac made available a new version of the CCM interface that included a batch search. This allows upload of one or more files to be processed offline. It incorporates the following functionality that takes into account the batch search description specified in the original project schedule together with the feedback from the project partners.

CCM Phase 3 Batch Interface Functionality v 27th May 2011
1. The ability to set up an RSS feed that will tell you when the results are available.
2. A Search History button on the search screen that lets you look at all your batch search results.
3. No limit on the number of records for batch searching.
4. Revised search procedure behind-the-scenes.
5. More information in the brief record.
6. A basic full record html display.
7. Records are sorted by author/title; records with no author will file at the top in title order.
8. Export of visualisation data now saves as a csv file.
9. There is a MARC exchange export format.

The project partners discussed and agreed the following contents of the MARC exchange record:

```
001 <Copac record number>
003 UkMaC
035 $a<searching libray's own record number>
245 $a<Title><statement of responsibility><publication details><edition>
[583 - potentially; but there are no examples yet]
959 $a$aCopies held by: <n> institutions$b<date stamp>$z<no. holdings institutions>$5UkMaC
```

Testers were alerted to the following:

a) In most cases there will only be a single 035 in the record. However, where Copac have consolidated two or more of a library's records within a single Copac record, the exported record will include multiple 035s. This will remain in the short term as alternative options

potentially take development into interface territory which will be better dealt with beyond the current limits of this project.

b) A search for a set of ISBNs will possibly result in some of the records in the result set export not containing a 035. This may be because the searching library doesn't hold the document in question, or because Copac has duplicate records for a particular ISBN.

c) To overcome the problem noted at the May project meeting regarding the csv export and Excel tending to strip 0's from ISBN's (and this can be a problem with the downloaded visualisation data), a temporary fix was applied. The export has been set up to put ISBN's in single quotes - that way it gets treated as text and the zero's are preserved.

To help the testing a few changes were made to the interface where it was practical within the timescales. Rather than making piecemeal changes it was decided that the interface will be dealt with as a separate package if the work is taken forward into a pilot stage. Thus a number of the interface related issues that arose at the May full project meeting have been left on one side to be incorporated into later developments. [These were articulated in section 6 of the Interim Report.]

A further version of the interface was released on the 24th June incorporating the following features:

CCM Phase 3 Interface Functionality v 24th June 2011
1. On the initial search screen a subject search has been added below the batch search.
2. The MARC export option is now at the top of the result screen so it is easier to see.
3. For a 'number' search there is a new graph in which the numbers you searched for are shown on the vertical axis, with the no. of holding libraries on the horizontal. (This format has its limitations as the size of the result set increases, so it will only be present for a result of less than 200 records.)
4. For all types of search, and all sizes of result set, there is a new 'summary' graph. The basic layout is number of records on the vertical axis and number of holding libraries on the horizontal. But the format varies slightly depending on the search.
5. From the visualisations page, if you select the option to export the 'graph data' this will give you the information used to create the new graphs.

The final version of the interface that completed the basic development work for the Phase 3 testing was made available to testers on Friday 1st July. Features now include two further options that had been requested by the project partners.

CCM Phase 3 Interface Functionality v 1st July 2011
1. Option to limit a search to a single library, currently Leeds, Manchester, Sheffield, & York
2. Option to limit a search to the Yorkshire region, i.e. Leeds, Sheffield and York.

5. Phase 3 CCM Tool Interface Testing

A testing framework had been agreed by the Partners earlier in the project as has been described in the Interim Report. (For completeness the testing framework table is reproduced in Appendix 1 of this final report.)

5.1   Testing Plan

To ensure consistency of approach and understanding across the test sites a testing plan for the new interface functionality was agreed based upon the two main criteria for the CCM tool:

- Criterion 1: Can I release this book?
- Criterion 2: How does my collection compare in strength to other UK Libraries?

The table in Appendix 2 details the actions for testers.

Leeds also did some further comparisons between subject areas e.g. comparing holdings between different modern language collections to see how they compare both nationally and within Leeds. York also looked at identifying subject areas of comparative strength and weakness.

5.2   The Test Reports

The areas to address in the Phase 3 testing were developed and identified in response to the Phase 2 work. It is worth reiterating here the point made in section 3.4 of the Interim Report:

"….that the test tool is a device to explore proof of concept and to specify required functionality is reflected in the commentary of the full test reports….." and like the Phase 2 activity the nature of suggestions for future development found later in this report, reflect this.

The Phase 3 interface testing was undertaken at Leeds, Sheffield and York Universities. Each site has produced its own separate report, and at a project meeting on 8th July the findings and observations were fully discussed. Due to lack of local technical systems support at the time of testing York was unable to complete on criterion one, or criterion two (part one) testing. The individual test reports are provided in the following sections of this report with method and observations being highlighted.

Overall Impression

In line with the overall impressions from the Phase 2 testing the general view was very positive. The addition of the batch functionality was thought to be an essential element for the tools success. One of the key observations agreed by all of the testers was that the subject term search can be used to quickly explore very specific strengths in the "national collection" and that a valuable feature of the tool is felt to be its relevance for collection development/management work in terms of identifying subject areas of comparative strength/weakness or identifying the level of occurrence of individual items across other libraries, (at a national or more local level) which would input to withdrawal decisions.

The opportunity of this seven month project has provided the means to make a significant step forward towards specifying and developing the core functionality for a tool that could contribute to a more coherent and sustainable approach to collection management. The view of the team was

that the conclusion of the Phase 3 testing has the potential to provide a springboard for taking forward findings into the development of a much needed pilot service.

The observations from the testing provide valuable input into refining functionality and will give confidence to a pilot specification and development if there is opportunity to take forward this work.

5.2.1 Report from Leeds

Criterion 1: Can I release this book?

Method - A batch file was submitted with 41 record control numbers. The output file was examined using the MarcEdit program. The file contained the following data fields:

- Record number with Copac library prefix
- Record number without Copac Library prefix
- Brief bibliographic description
- Holdings info:
  - Number of copies held nationally
  - Date of search
  - Number of copies held nationally in numeric format
  - Mimas identifier

This is the data which had been agreed for the initial testing. The MARC record was valid and was successfully processed by the Innovative LMS at Leeds. With a little effort to create the correct loading profile, national holdings data could have been loaded into each of the records (matching on control number) and used as part of a programmatic and automated discard process provided a final human check was included. Only one member of Leeds staff is able to create loading profiles, so it was not possible to test this final stage in the live system. However it is clear that the data is adequate for this purpose.

Observation: An option to export these same data fields in comma-delimited format for loading into Excel or Access would be useful for stock management processes which are independent of the Library's LMS.

The same batch file of control numbers was submitted, but limited results to "Yorkshire region", with a view to consortial stock management. The data visualisation options were explored and the data output in MARC format. It was noted that in both cases, although the set of records was restricted to those with holdings in the Yorkshire region, all holding libraries were included in the visualisation and similarly in calculating the number of holdings in the MARC 959.

Observation: Both visualisation and MARC export should reflect only holdings in the "Yorkshire region" i.e. the local consortium. However, it was realised that for the MARC export, it would be important to flag that the total number of copies stated was for the consortium and not national, so an additional subfield is required which clarifies this and can be stored with the holdings data.

Within this context, it would be worth exploring whether inclusion of the individual holding libraries might provide useful data for decision making, since this information would be more meaningful within the more restricted environment of a local consortium where collection management policies, etc. of the partner libraries are likely to be better known.

Changes to the online interface were tested.   Firstly, data presentation in initial list of hits against the database (sort order, fields present, etc.)

Observation: The sort order is much improved in that it generally brings together records for the same bibliographic item. However, it is not easy to locate a known item because of the use of the author field in the sorting algorithm. It is suggested that a straight title sort would be more helpful. Other fields are adequate for a brief display, and the link to the full record is very helpful. It would be useful to be able to browse from one full record to the next as is possible in the main Copac interface.

Secondly, data visualisation tools (esp. export of holding libraries as csv file)

Observation: The mouse-over in the graphs to give more detail is extremely useful but in the Number of records by library graph, it would be helpful to also show the name of the Library as this becomes very unclear if there are a lot of holding libraries.

Export of both map data (count of holdings per library) and underlying graph data (count of holding libraries per record) is very easy.

Once the visualisation option has been chosen, there is currently no way to return to the original list, nor can you right click the "Visualisation" button to open in a new tab.

Can the search for a single item be embedded in other applications?

Observation: When viewing a specific record within the Leeds "Millennium" Library Management System, it was possible to launch a search of the tool based on that specific record.

An option was created in the Millennium LMS which allows the record displayed to be searched against the Copac CCM tool and the result displayed which could be used to make decisions (e.g. to conserve, to discard, etc.) about individual items as in use cases 3 & 4.

Observation: For speed of development, ISBN lookup was implemented rather than record number, but the latter would be a possibility. It would be possible to modify this facility to show regional rather than national holdings if desired. This is an instance where the map visualisation is potentially very helpful.

The ability to limit results to specific libraries in the online interface was tested, using both ISBN and record number to search for a single record.

Observation: As with the batch functionality, it would be more helpful for the results and visualisation to reflect only those holdings in the subset selected.

Within this context, only a few holdings are displayed, so the holdings list in the primary listing is easily interpreted and therefore the visualisation provides less added value. However, in contexts where larger result sets are generated, this balance would change.

Criterion 2: How does my collection compare in strength to other UK Libraries?

Method – Two large batches of ISBNs (representing ca. 4,000 records from each of modern French literature and modern German literature) were submitted. [See the detailed case study in section 7.]

Observation: It was noted that this gave a good picture of the overlap with collections in other Copac libraries (as in phase 2). It had been hoped that the graph showing how many records were held by just one library, by two libraries, etc. would allow conclusions to be drawn about the number of unique items: - since all items are held by Leeds, this would equate to unique to Leeds, Leeds and one library, etc. However, because of record duplication and because the graph did not amalgamate the holdings attached to several records which related to the same submitted ISBN, this second dimension was not able to be drawn out from the data as presented. If the functionality is updated to support this, then it seems likely that an accurate profile of the collections could be drawn up.

The graph displayed for small sets showing how many libraries hold each title seems less useful, though the actual implementation with the mouse-over giving the bibliographic details and the holdings is very neat.

The ability to download the "graph" data as well as the "map" data is extremely useful, as is the fact that the resulting file names are linked. This can be used to recreate locally the "number of records" graph, albeit with a little effort. Consideration should be given to providing a downloadable version of that graph.

Some work has been previously been undertaken nationally on collection overlap. Consideration should be given to providing functionality for an option to explore this via these result sets.

Comparison of result sets for different languages was hampered by the lack of specific information from the "number of records" graph, so this was not systematically investigated. It can be noted that the degree of overlap for our modern German holdings seems to suggest collections at least as good as for modern French, which is useful information. What we don't know is how much overlap can be expected with language collections (where different editions and a variety of works by the same author tend to increase the number of unique holdings compared to most other subject collections).

Subject searching.

Some testing was carried out on subject searching and correlating this with number searches based on specific subject areas. Results are summarised below.

A number of LCSH terms were used, all relating to Jewish liturgy.
- Judaism – Liturgy – Texts
- Seder
- Pilgrim festivals (Judaism)
- Haggadah

Observations: *It was noted that these searches resulted in more hits for Leeds in Copac CCM than in our own catalogue, and it seems that the Copac search is by keyword rather than a subject term search. Given the need to be very precise to avoid large and unmanageable result sets, and also given the way in which this search is presented, it would seem more helpful for it to be based on a phrase search for the precise subject term.*

For these subject terms, the number of records by library graph showed a small number of libraries which had strong holdings for all these subjects: British Library (by far the strongest), then Cambridge, then either Leeds or Oxford. This in itself gives a good feel for the geographic distribution – size of balloon on the map would be helpful here.

A set of 6,417 records from a specialist collection of Judaica held at Leeds was then submitted. The 7 libraries most represented were Cambridge (661), UCL (661), Oxford (626), Manchester (584), Southampton (503), TCD (392), British Library (371). The low figure for the British Library probably arises from the nature of their records, which makes more difficult record clustering for the same bibliographic item. Indeed, the low overlap overall arises from the same cause and illustrates the need to include some form of de-duplication algorithm as an option within the CCM Toolkit. The scope of this collection was much broader than liturgy, which is probably why UCL, Southampton and TCD show comparatively strong holdings. Equally, it seems likely that the British Library, Cambridge, Oxford and Manchester are clearly strong in this area alongside Leeds.

Observation: *This example shows how the subject term search can be used to quickly explore very specific strengths in the "national collection". Here too, size of balloons on the map visualisation would be helpful.*

5.2.2 Report from Sheffield

Criterion 1: Can I release this book?

Method - Search by entering a comma delimited set of standard numbers:

- Submitted a comma delimited list of 20 local record numbers (the same numbers as used in the Phase 2 test to aid comparison).

Search results: 20 records were found by the search, as for Phase 2, but there were some differences in the number of libraries found. Five of the 20 records were different, some having less libraries, some more. Overall, an extra 21 libraries were found, as compared with Phase 2, bringing the total from 162 in Phase 2 to 183 libraries in Phase 3 - more in line with the ISBN search results in Phase 2.

MARC export of results: Not able to do this. Will need assistance from the System and Database Section of the University of Sheffield Library.

Visualisation of search results: Three graphs all useful in their own ways. Axes clearly labelled in most cases. Some axes partly labelled (due to lack of space for the information required on the axis?), which makes identification more difficult. Use of different colours for bars on the graphs very

helpful, as is the pop up information that displays when the mouse cursor is placed over a bar of the graph.

Observation: As well as the number of locations being displayed, would it be possible for the names of the locations to be listed? This would also help in clarifying the graph where not all of the locations are noted on the axis due to lack of space mentioned above.

Map useful for showing geographical distribution of search results. Ability to zoom in and out of map very useful, as is the pop up information that displays when the mouse cursor is placed over a location marker on the map. A possible improvement would be to have varying sizes of marker according to the relative numbers of the search results.

Observation: As for Phase 2 testing, the graphs and map were still not easy to print out. Could this be improved on?

Export graph data facility very quick, easy to use and very useful for stock editing purposes. The csv file produced can easily be converted to an excel spreadsheet and then worked on e.g. if a certain number of locations was used as a benchmark for potential withdrawal, say, eight or more, then the data could be filtered to exclude all those rows where the number of locations is seven or less, and you would be left with a list of items for potential withdrawal.

Observation: Export map data facility useful if work being done where the geographical information is important e.g. where looking for regional distribution of items. The csv file produced, as with the export graph facility, can be easily converted to an excel spread sheet, and worked on effectively.

Method: Submitted a comma delimited list of 20 ISBNs (the same as used in the Phase 2 test to aid comparison).

Search results: 44 records were found by the search, as compared with 45 in Phase 2, and there were some differences in the number of libraries found. Six of the 20 ISBN records showed different locations, two showed similar locations but included different ISBNs than those inputted, and 12 had the same locations but six of them had split records in the result display. There seemed to be no logical order to the ISBN results, but it was noted that more of the items with the same ISBN were together, than for the Phase 2 testing. Overall, an extra 16 libraries were found, as compared with Phase 2, bringing the total from 185 in Phase 2 to 201 libraries in Phase 3, more in line with the WRBL CCM database search results referred to in the Phase 2 testing.

MARC export of results: As above

Visualisation of search results: As above

Method: Search by uploading a file of standard numbers: Not able to do this. Will need assistance from System and Database Section of the University of Sheffield Library.

Criterion 2: How does my collection compare in strength to other UK Libraries?

Method: Search by subject

Fascism was chosen as a subject as the University of Sheffield has a Fascism Collection and it was not known how unique or common the items were nationally.

- The first search was 'fascism' limited to 'the University of Sheffield'.

Search results: 897 records were found, which would seem accurate, as there are 532 items in the Fascism Collection and there will be other items on fascism in the other collections of the University Library.

Visualisation: The graph showing number of records held by number of libraries, showed that 182 of the items were unique to Sheffield, and that the other 715 items were held at between two and 31 libraries. This graph quickly and effectively shows how unique or common the items in the search are across the Copac libraries. The graph showing the number of records by individual libraries showed that the library after Sheffield holding most of the items was the British Library (509), then Cambridge (436), then LSE (401).This graph also quickly and effectively shows how much or little of the items in the search are held in which particular libraries, which could be very useful for comparison of collections, for sharing resources, for weeding and building up collections etc. The map showed the geographical distribution of the holding libraries of the fascism items very well, but as mentioned above does not give any indication of the comparative size of holdings, which would be helpful.

Method: The second search was 'fascism' limited to 'the Yorkshire Region'.

Search results: 1481 records were found, which seems reasonable as it includes records from the Yorkshire region rather than just the University of Sheffield.

Visualisation: The graph showing number of records held by number of libraries, showed that 182 of the items were unique to Sheffield (as expected from the previous search), and that the other 1299 items were held at between two and 31 libraries. Although the graph quickly and effectively shows how unique or common the items in the search at the University of Sheffield are across the Copac libraries, it does not show how unique or common the items of the Yorkshire region are across the Copac libraries. The graph showing the number of records by individual libraries showed that Sheffield with 897still held the most, then came the British Library (842), then Leeds (784), then Cambridge (723).This graph also quickly and effectively shows how much or little of the items in the search are held in which particular libraries, which could be very useful for comparison of collections, for sharing resources, for weeding and building up collections etc. The map showed the geographical distribution of the holding libraries of the fascism items very well, but as mentioned above does not give any indication of the comparative size of holdings, which would be helpful.

Method: The third search was 'fascism' limited to 'all libraries'

Search results: The search took a few minutes, but eventually 10971 records found, which seems reasonable as it includes records from all libraries rather than just the University of Sheffield.

Visualisation: The graph showing number of records held by number of libraries, showed that 182 of the items were unique to Sheffield (as for the previous searches), and that the other 10789 items

were held at between two and 26 libraries. The graph quickly and effectively shows how unique or common the items in the search at the University of Sheffield are across the Copac libraries, which is very useful for us at the University of Sheffield. The graph showing the number of records by individual libraries showed that Oxford with 3496 held the most, then came the British Library (3214), then Cambridge (2613), then LSE (2611).This graph again quickly and effectively shows how much or little of the items in the search are held in which particular libraries, which could be very useful for comparison of collections, for sharing resources, for weeding and building up collections etc. The map showed the geographical distribution of the holding libraries of the fascism items very well, but as mentioned above does not give any indication of the comparative size of holdings, which would be helpful.

5.2.3 Report from York

Criterion 1: Can I release this book?

Technical systems support was unavailable to input to the testing due to other Library priorities, and therefore no testing was completed on criterion one, or criterion two (part one.)

Criterion 2: How does my collection compare in strength to other UK Libraries?

The following observations from the testing were made available:

a. York testing focused on the testing of subject searches and the evaluation of results. Limited time was available for testing on this occasion.

b. The value of this feature of the tool is felt to be most relevant for collection development/ management work, in terms of identifying subject areas of comparative strength/weakness or identifying the level of occurrence of individual items across other libraries, (at a national or more local level) which would input to withdrawal decisions.

c. Subject search terms: Able to enter a combination of terms. Not sure if there are limitations to the number of terms entered?

d. Would be useful to be able to do an author search (i.e. tested on Poetry Collection – would have been useful to be able to search on author as subject to compare strength of holdings.)

e. Speed of search was acceptable and caused no major issues (maximum number of results produced 6000.)

f. List produced by subject search terms produced in alphabetical order, do duplicated records sit together – which is more useful than previously.

g. Graph showing "number of holding libraries" particularly useful, and the "explosion" of the individual bar to identify specific holding libraries: useful for interrogation of individual items for disposal.

h. Graph showing "number of records": useful for highlighting general strengths in comparison with other libraries – in particular comparing against like libraries (and in this situation, disregarding copyright/legal deposit libraries.)

i. Exported graph data: Did not inform beyond the information illustrated in graphs.

j. Map: of more limited use – others may find it more useful. Indication of comparative size of holdings would be useful.

k. Exported map data: Of more use. Can sort list to rank number of records and name of library from batch search, and assess own and others collection strengths/weakness.

l. In addition to use for identifying stock for withdrawal, highlighting collection strengths, informing researches/librarians of stock availability, would this have any use for Inter Lending services?

6. Use Case Scenarios

6.1 Interim Report Use Cases

The Interim Project Report described four Use Case scenarios that the project team had developed in light of the participating library's collection management requirements and experience of the first CCM interface trial, alongside the discussion arising from that work. These four covered the following:

Use case 1: Identifying last copies among titles considered for withdrawal

Use case 2: Identifying collection strengths

Use case 3: Deciding whether to conserve a book

Use case 4: Reviewing a collection at the shelves

Resulting from continued consideration of the benefits the CCM Tool offers to libraries and how easily it can be integrated into existing workflows two further Use Case scenarios have been developed during Phase 3 of the project. The purpose of these illustrative use cases will be to demonstrate how this tool can be used productively. The aim will be to look to applying and testing all the use cases in live conditions once the CCM Tool is more fully developed.

6.2 Detail of additional Use Case Scenarios

6.2.1 Use Case 5: Prioritising a collection or item(s) for digitisation

Context

Many libraries are faced with the monumental task of potentially digitising entire collections either through user request or preservation needs. Dependent upon resources and funding some material may be lost due to the lack of data required to understand what should be prioritised over other content. With all the competing priorities of a digitisation service (on-demand, research led, project funded, preservation) having a tool that identifies what is a strength, is unique or endangered speeds up the selection process and places content in the queue for conversion.

Procedure

The decision to digitise would be based upon the criteria below:
Submit records related to French studies (example)

- Run batches of records related to sub-disciplines within the broad area of French studies
- Based upon results it is determined that 18th century poetry contains a particularly large number of titles which are either unique or held by only one other library
- If multiple copies are held are there unique elements in the holding institutions copy that justifies digitisation?
- Cross check proposed items against Google and Europeana for digitised copies

- Collection(s) or item(s) can also be marked for digitisation dependent upon results of use case 2 and 3
- Digitised collection(s) item(s) are recorded in the catalogue or external holdings are referenced (recommendation for future reports and best practice)
- Once quantities are determined, digitisation costs can be estimated in addition to description and storage for item(s) collection(s).

Benefits

- Using tools such as the CCM tool to understand holdings, both in terms of individual items which are rare or unique; and whole collections which contain a large number of such items; reduces and possibly prevents duplication of effort related to digitisation.
- Research led, on-demand, and project based scanning represent significant strands of work within a digitisation service. Parallel to all this would be collection digitisation for access and preservation (preferably research led), which this tool would facilitate.
- The tool would support decision-making and prioritisation around digitisation, allowing better use of limited resources and enabling rare or unique materials to be slotted into the digitisation programme and workflow and prevent loss of irreplaceable research material.
- Understanding the collection strengths enables funding bids to be targeted toward those collections.
- Materials that are at risk can be preserved and made available or dark archived dependent upon copyright status, ensuring on-going access to them and contributing towards a national distributed resource for future generations of researchers.

6.2.2   Use Case 6: Subject search - Collection development and marketing

Context

Universities are under pressure to attract more students and high quality researchers. This often involves creating new courses and/or research areas and expanding the topics covered by existing courses or research. It also requires strong marketing. An example of the former would be creating a new department. An example of the latter would be a department which historically focuses on their topic in the context of Europe and the US, but in order to offer a competitive course they expand this to include Asia and the Pacific, or even the whole world. The library's collections will often be weak in the new area and will need building up. Libraries also need to market what they already have more effectively.

Procedure

- A department is expanding its subject coverage to include art from Asia. The Library needs to assess comparative strengths and weaknesses in relation to comparators/competitors and to identify the resources held by libraries which have a strong collection to inform a collection development and budget plan.
- The Library does a subject search on the Copac CCM tool and reviews the data produced.
- The graph showing number of records indicates which libraries have strengths in the subject area. The graph showing number of holding libraries indicates which titles are core to the subject area and which might be considered for purchase.

- The results are exported into a spreadsheet to allow more detailed analysis of what constitutes a strong collection and this forms the basis of a collection development plan for discussion with academic staff backed by evidence. Legal deposit libraries can be excluded or results confined to a subset of libraries (e.g. regional or competitors in the subject area as identified by academic staff).
- The evidence is used to support bids for additional start-up funding and for prioritising purchases.
- The process can also be used to support decision-making in filling known collection gaps.

Benefits

Without the Copac CCM tools there is no objective way of comparing the strengths of libraries' collections or identifying which are the core resources in a subject area. The analysis enables evidence-based collection development plans, budget bids and prioritisation of purchases. It also opens the way for negotiation with other libraries, including potential transfer of collections. It can also be used as a marketing tool. The evidence can demonstrate when collections are strong and this can be used to attract students and academic staff.

7.   A detailed case study - Identifying collection strengths

The testing described in both the Interim Report and this report has led to very interesting demonstration of the value of applying the CCM Tool to real life challenges experienced by library staff at Leeds University. This section details some of that work and key findings, only made possible through this project, and it also illustrates where there is still work to be done on refining the tool to enhance its value to collections management.

Use case 2 (*Identifying collection strengths)* is based on a programme of work to which Leeds has been committed for over two years. In order to assist prioritisation, decision-making and resource allocation, Leeds is currently trying to identify those collections which it believes will be of particular value and significance in the long term, and which can therefore be designated as "heritage collections". These generally represent collections which have national or international significance for the research community and should be considered key constituents of the "National Distributed Research Collection" advocated in various reports and documents produced by the RIN and others. Initially, this work at Leeds was based primarily on subjective assessments by staff familiar with the collections in question. Latterly the Copac Collection Management Tools Project has provided an opportunity to explore the possibility of basing this work on statistical evidence drawn from the Copac database. This case study uses two actual sets of data to illustrate how the outputs can be used and the limitations inherent in the current stage of development of the tools.

Within the Leeds framework for categorisation of collections, one criterion which will define a particular collection as "heritage" is that it of "national or international significance" and it is suggested that evidence suggestive of this might include:

- That the collection contains a high concentration of items which are rare or unique (i.e. few copies are recorded in Copac),

And/or

- That the collection is particularly extensive compared with the holdings of other UK libraries for the same subject area.

Staff at Leeds were aware that the University's interest in Colour Chemistry, dating back to a Department of Dyeing in the original 19th century Yorkshire College, meant that holdings were particularly rich within this discipline. It was therefore decided to use the Copac tools to compare a subset of the 21st century acquisitions in Colour Chemistry with a similar subset for Chemistry in general to see whether any differences emerged. ISBN searching was used to accommodate the existence of duplicate records for the same title within the Copac database. Figures 1 and 2 below shows the national holdings for these two sets of records and reveal a clear difference. For Colour Chemistry, Leeds is shown to hold significantly more items than any other library except the British Library, so this collection is indeed exceptional (even the legal deposit libraries only seem to manage 75% of the Leeds holdings). In contrast, although the holdings for Chemistry are also strong, Leeds ranks only 4th nationally, after BL, NLS and TCD, and with Cambridge, Oxford and NLW not far behind. The two graphs demonstrate this difference in a striking visual manner and suggest that it might be possible to describe in quite precise and mathematical terms the difference between a heritage collection and one which is merely "strong".
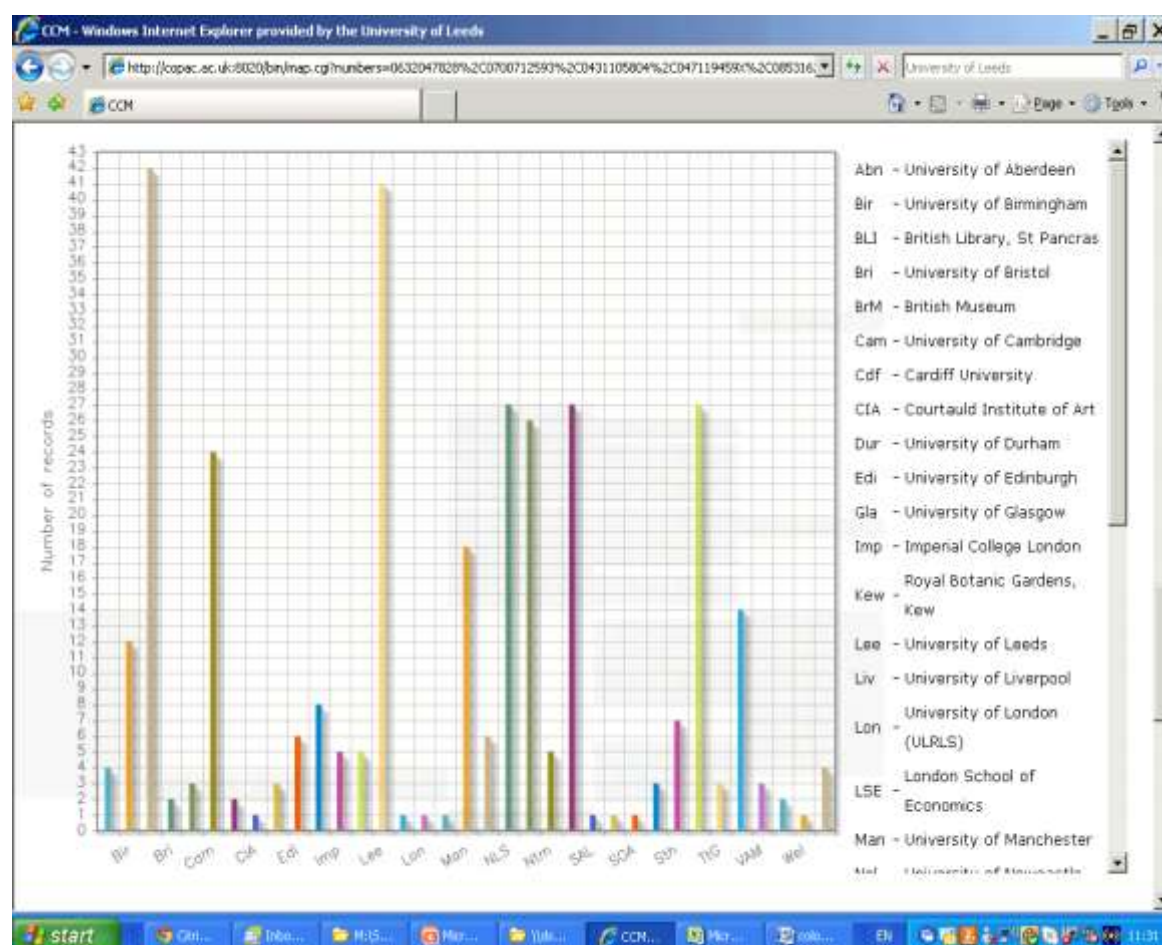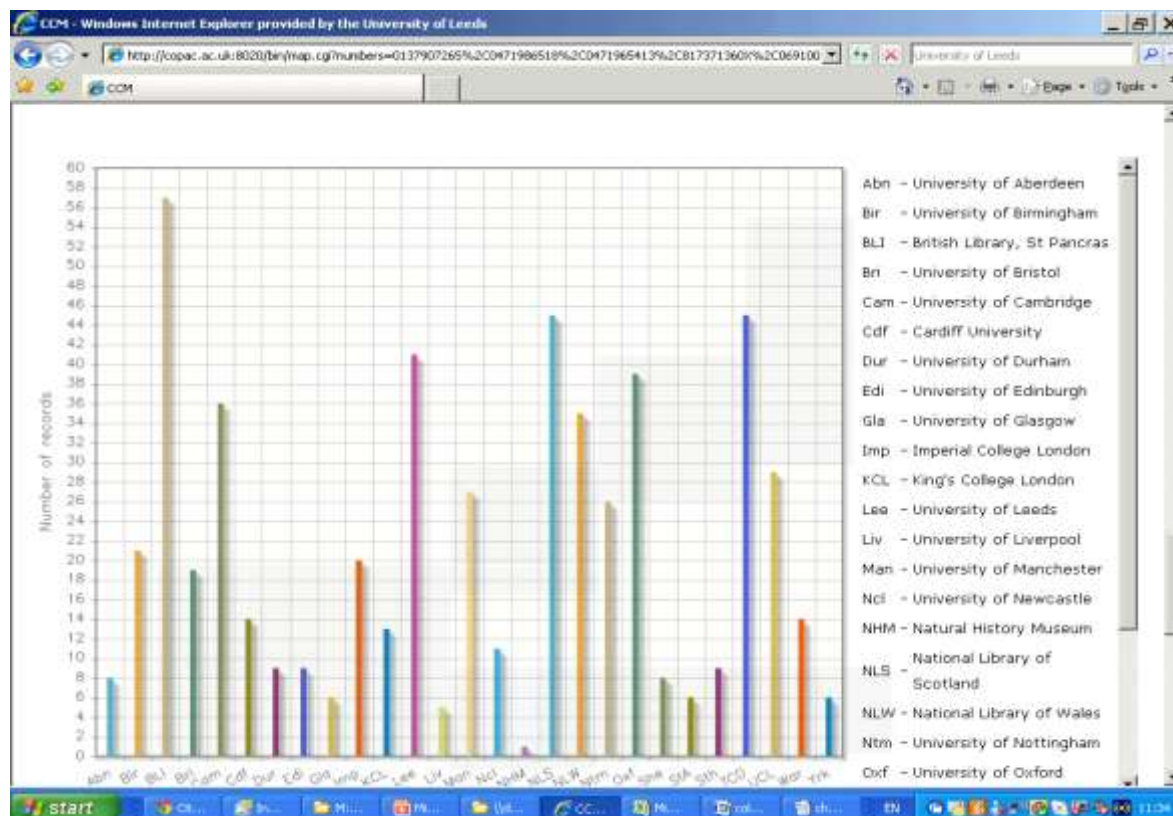
Figure 1: Colour chemistry – national holdings

Figure 2: Chemistry – national holdings

In the final phase of the project, during which larger subsets of data could be analysed in batch mode, Leeds decided to compare its holdings of post war literature in French and German. Both are important departments but with slightly different histories in terms of collection development and it would be useful to know whether either collection showed clear evidence of "heritage" status. It was also hypothesised that for literature subject areas, there might be less overlap between individual university libraries, since the same literary work can be available in a number of bibliographically distinct editions and the same author can be represented by a different selection of individual titles. If true, this would imply that the graphs for literary collections such as French and German would be somewhat different from those for scientific disciplines such as Chemistry or Colour Chemistry, even where they are both of national significance and categorised as "heritage collections".

Figures 3 and 4 below show the national holdings for the Leeds collections of modern French and German (again based on ISBN searches).

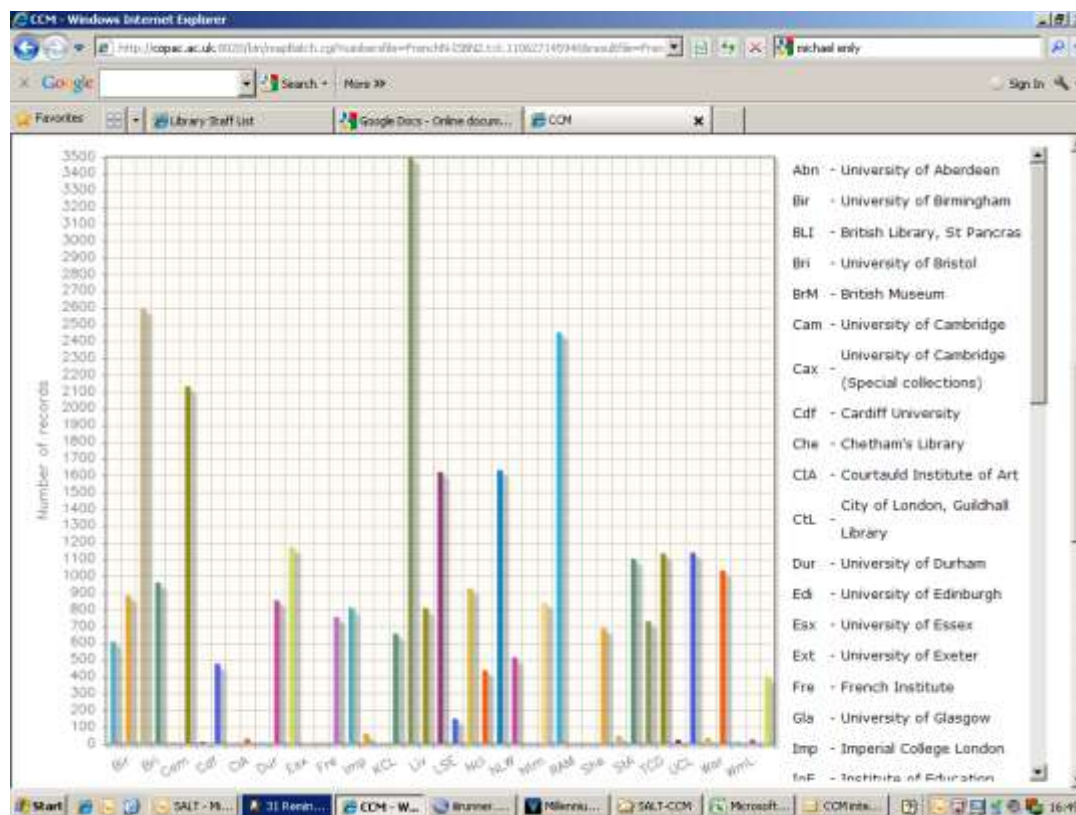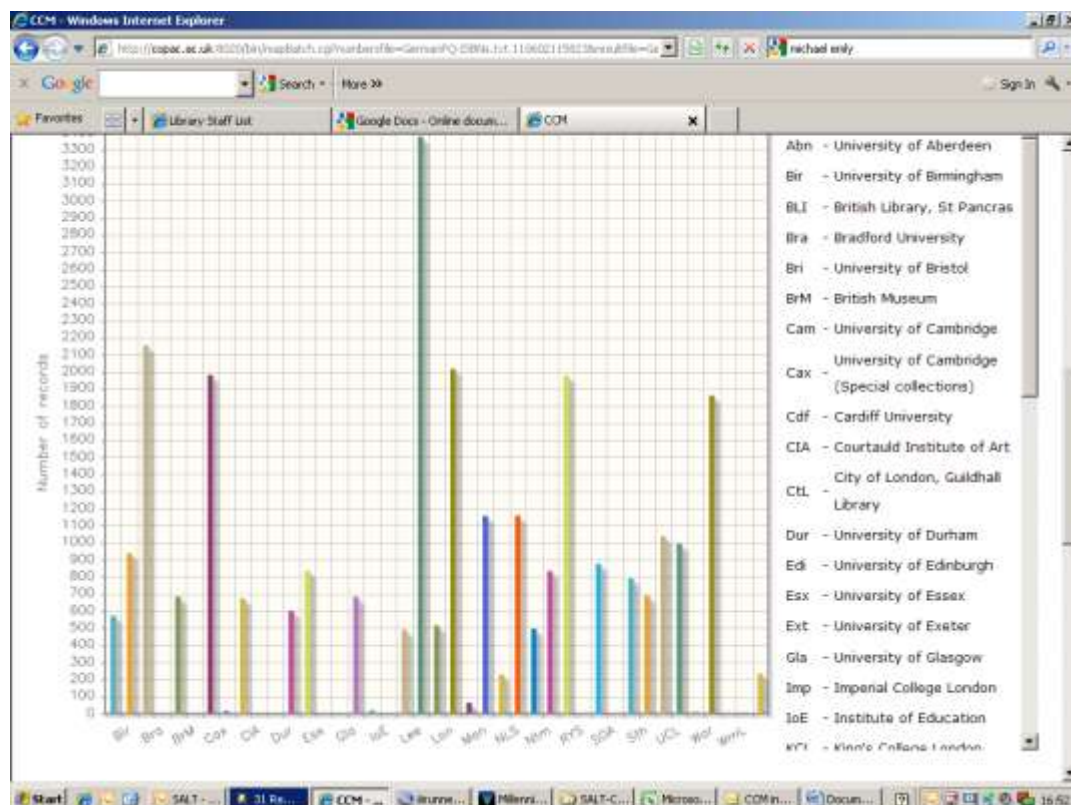**Figure 3: Modern French – national holdings**

Figure 4: Modern German – national holdings

These do seem to suggest a different pattern from those in figures 1 and 2 for scientific disciplines, and it is very noticeable that the British Library does not seem to hold such a high percentage of the editions held at Leeds.   It is also interesting to note the internal differences between French and German, both in terms of the individual libraries which seem to have strong holdings (Warwick, for example, seems particularly strong for German – or perhaps it is true to say that its holdings mirror more closely those of Leeds) and also in the overall distribution of titles. Whereas these graphs show which libraries hold a high proportion of the titles held at Leeds, it is also important to ask how many of the Leeds titles are unique to Leeds or not widely held by other UK libraries. The Copac tools provide a graph intended to answer this question (figures 5 and 6), but a key obstacle immediately becomes apparent: because of the existence of duplicate records within the database, the vast majority of records are actually shown as held by just one library, and that one library not Leeds, whereas we know that this is not in fact the case since all the ISBNs originated in Leeds. Algorithms to identify and merge such duplicate records will be applied as part of the forthcoming migration of the Copac database to a new platform, but it would also be desirable to further develop the tools themselves so that record "merging" can be carried out "on the fly" at the reporting stage, based e.g. on a shared ISBN, as is currently achieved for the graphs and maps showing national holdings for the set as a whole. Meanwhile, it is not possible to identify what percentage of these Leeds collections are unique or rare.

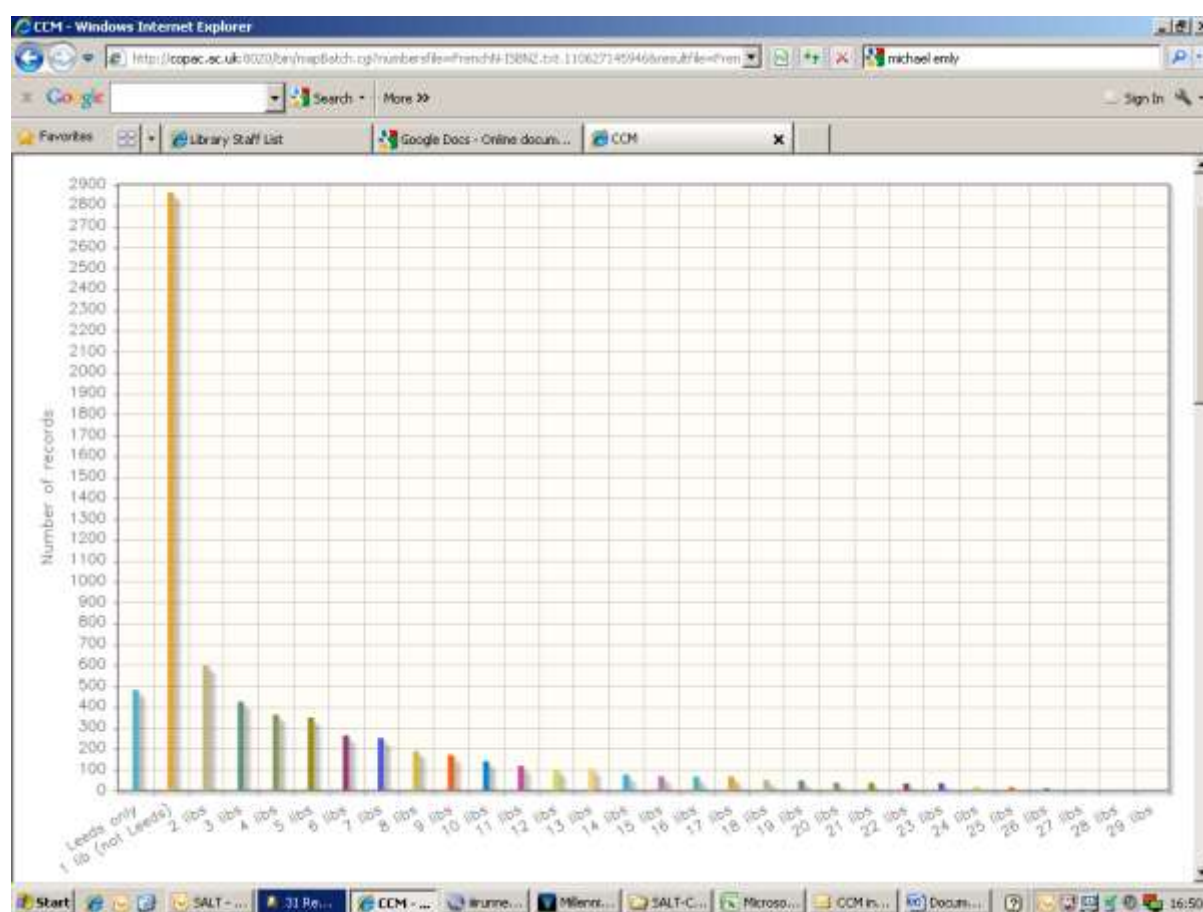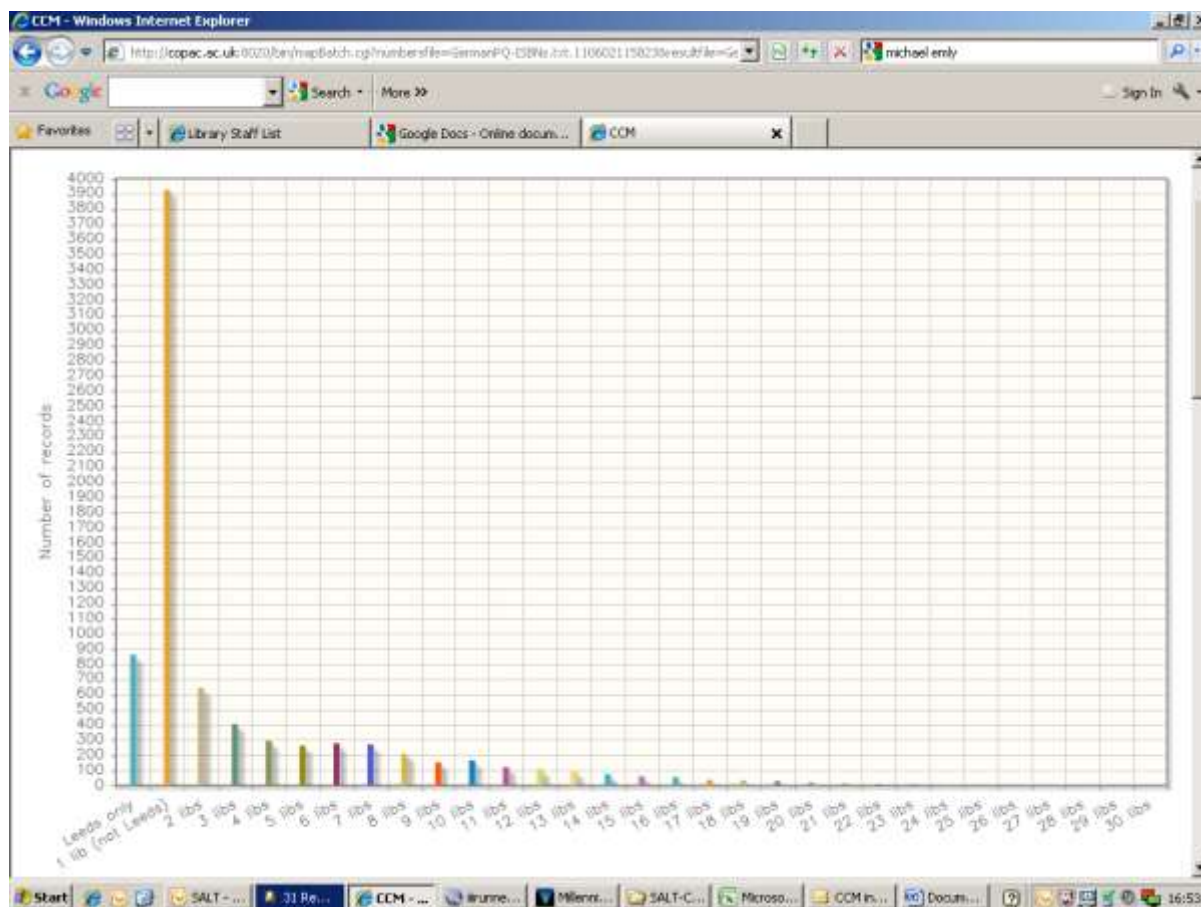**Figure 5: Modern French – unique and rare titles**

Figure 6: Modern German – unique and rare titles

This case study, based on the current version of the Copac tools, demonstrates both the need for further development and refinement of those tools, but also their potential for answering strategic questions about the status of the collections in individual libraries and the opportunity to deepen our understanding of the parameters around collection development in terms of:

- Overlap between the holdings of major UK research libraries in particular subject areas;

- Differences in that overlap between different subject disciplines and areas;

- The proportion of unique titles within those collections;

- The extent to which researchers will find that they no longer have access to such a wide range of research materials in future years, given current pressures on space and the widespread severe deterioration of printed materials through brittle paper and collapsing bindings.

8. Current Status of the Project

With the completion of the Phase 3 testing there was a desire amongst the libraries to continue work on developing case studies and use cases. Thus some further development work was agreed to continue beyond the end of the project as resource permits. Notably, an optional ISBN de-duplication of result sets, to reduce the impact of record duplication on the visualisation of results; an author search option; regional filtering expanded to include the visualisation displays; and mouse highlighting line on graph to bring up name of institution.

One area outstanding from Phase 3 and still requiring further investigation is the task to agree a scale of retention levels with definitions of those levels and a discussion of the potential for limiting results by retention status. This is a complex area. The Leeds partners are intending to take forward discussions with the British Library on this topic over the summer months. The project partners have discussed whether it would it be possible for Copac to take the initiative of "marking records as only copy" as a starting point for libraries. There are questions of how to maintain this information accurately when Copac get updates. 'Pushing' this information to interested libraries would be a possibility. But this is an area that would require further investigation.

The CCM interface will remain available to the project partners until the outcome of a proposal for a continuation of the project is known.

The project team intend to disseminate the outputs from the project, including the Final Report via the Copac and RLUK blog, and announce its availability on lis-link and the JISC Collections Management list. An article based upon a use case is also to be explored along with an appropriate avenue for publication.

9. Future Potential and Recommendations

The Interim Report identified within section 6 "Forward Look", the desire of the project partners to "*build upon the work of the project beyond the current limits and to develop it further. The aim, as stated within the original project proposal to the JISC would be to roll out a pilot service, in the first instance to RLUK libraries to act as first movers, applying the tool within their workflows.*"

The development areas for future attention specified in the Interim Report covered: the interface; search expansion and post-search de-duplication; retention status; and collaborative support. In addition to these, at the end of the Phase 3 activity, the following areas need also to be considered in any future extension to the CCM Tools Project:

- Expansion of test libraries and resilience testing of the tool. Address evidence of scalability.
- Building collaborations and alliances with interested organisations pursuing complimentary activity
- Addressing the development of a business model for a service beyond a pilot
- The development of a detailed communications plan ensuring dissemination of information and engagement at all levels, across stakeholders and interested parties.
- Dissemination to include a workshop with collection managers based on the current project and use cases. (Optimally this would be around November or December 2011.)

The six use cases described in the Interim and Final Reports articulate the practical benefits that the CCM Tool could bring to supporting collection management in this country. By leveraging this potential within Copac, an existing community database, this prototype project has demonstrated that there is indeed a need for such a tool, and a solution, to help universities make informed decisions about monograph retention and disposal and collection development. That Copac is ideally suited to this reuse of data has been demonstrated through the outputs and observations of this project.

Appendix 1 – Testing Framework

The project has agreed core criteria for the testing framework and areas for development. This work has been summarised in the table below. Potential methods for testing and assessment, forming the basis for the Testing Plans for Phases 2 and 3 have been derived from this analysis.

Criteria	Comment	WRC Notes	Assessment
1. Can I release this book?	*Noted risk – currency of content because of variable frequency of library updates. May be mediated by access to live circulation data.* *Noted risk – Split location & item record likely to cause false results; especially for items which initially appear to be very rare. Is suggested that accuracy will be improved in the 'new' Copac with the major rebuilding of the database, though this is unlikely to be available within the timeline of this Project.* *Noted risk – Confidence in distinguishing between different editions, imprints and bindings of monographs may be of varied importance to different users in stock editing. This would be assisted by the provision of variable 'fuzzy matching' of accuracy which could be modified directly by end-users – possible approaches to be explored.*	Does the tool allow easy-to-use, clear identification of numbers and distribution of duplicated copies of monographs held elsewhere in the UK (to support localised withdrawal of 'zero use' items with assumption that sufficient copies remain to support future research needs). Should ideally allow for comparison of all Copac member's holdings, or for sub-sets and highly customisable searches (e.g. of particular institutional collections) by end-users. Criteria 1 is <u>arguably</u> more likely to be used by administrative/ clerical staff in editing procedures, as such tool interface and results should be particularly robust and clear.	See Below for detail. General issues for assessment are gaining an end-users impression of the clarity of interface, availability and easy use of options (e.g. more advanced customisable searches), speed and accuracy of responses, interpretation and manipulation of varied formats of results, and the ability to download data and link these results back to local LMS for local editing and record annotation. Compare accuracy of results with those from similar search in previous WRC work. Would also seek feedback from WRC institutions on how the Copac Tool could possibly support our users (i.e. students, researchers, academics) and what search options, and results would be particularly useful, developing some use cases. Of major importance in Criteria 1 is tying the results of the search back to the individual copy which was submitted by the user, this for marking items for potential discard or retention (of particular importance for batch search process). However, role and functioning of "fuzzy matching" will also be evaluated.
2. How does my collection compare in strength to other UK Libraries?	*Noted risk – Confidence in distinguishing between different editions, imprints and bindings of monographs may be of varied importance to different users in stock editing. This would be assisted by the provision of variable 'fuzzy matching' of accuracy which could be modified directly by end-users.*	Does the tool allow easy-to-use, clear identification of the numbers and distribution of duplicated copies of <u>groups</u> of monographs held elsewhere in the UK. Should ideally allow for comparison of all Copac member's holdings, or for sub-sets and highly customisable searches (e.g. of particular institutional collections) by end-users. Criteria 2 is <u>arguably</u> more likely to be used by specialist staff, the tool interface and results should therefore support a more flexible approach to use and offer more powerful customisation and display options.	As Above, though particular attention will be paid to users ease of interpretation of the results of 'collection strength' searches. Is the potential for confusion and/or misinterpretation of results when comparing duplication levels for a number of records across a number of institutions? Of 'potential' importance in Criteria 2 is tying the results of the search back to the individual copies which were submitted by the user, this for marking items which may be unusually rare (or common) elsewhere in the UK. The potential of different types of search and, again, the role and functioning of "fuzzy searching" to be evaluated.
Areas to look at	**Comment**	**WRC Notes**	**Assessment**
1. Interface			

Areas to look at	Comment	WRC Notes	Assessment
Search limit – currently 100 numbers	Not very meaningful for criteria 2, but OK for 1.	<u>Criteria 1</u> a) Individual items – quick search. <u>Criteria 1 and 2</u> b) Small groups of items (up to maximum of c. 100) – quick search.	**Assessment for all areas noted below is primarily by end-user testing of the tool, and comparing impressions of the process –and its applications - with colleagues (and also comparing experiences between the three WRC institutions).**
Batch size limit of 400 – 4000 more useful	To be handled in batch search process.	<u>Criteria 1 and 2</u> Might be useful to test larger batches at specific times?	As Above
Results display	Need to investigate the potential of the link to csv data for locations and any further developments in that area *(see also 6)* Could be useful to group/summarise by search term. Perhaps particularly important for "fuzzy searching"?	Access to brief key bibliographic details 'on screen' and online link to full record for non-batch search.	As Above
2. Map & graph			
Visualisation of results How is it useful? How could it be improved?	Map less important for CCM criteria 2; potentially good for end-users.	<u>Criteria 1</u> Main value of map may be for identifying closest 'local' duplicated copy for researcher access. <u>Criteria 2</u> Clarity of results is critical e.g. what does the map show (single copy from a batch or multiple copies at a single site). Customisable nature of search important to assist in this interpretation. ALSO how does main batch search interact with map display?	Seek feedback from WRC institutions on how the Copac Tool could fit into procedures for identifying collection strengths, and sorts of use this information may have on internal collection management (and on distributed 'national collection' management).
3. Can we action on conservation and preservation?	Preservation status - transaction code (583). Experience of scale of this. Identification of legal deposit items?		Seek feedback from WRC institutions on how the Copac Tool could fit into procedures for local prioritisation for conservation, and what information would be useful.
4. Workflow and use cases	Web interface useful for one-off discards and also donations. Longer term: create mobile version that could be used at the shelves with barcode scanner giving ISBN to initiate the search. Could be useful to generate a variety of use cases for different aspects of the system.	ISBN option useful, but is anticipated that many non-ISBN items (pre 1970s) will be searched, so mobile device must allow other search inputs.	Seek feedback from WRC institutions on how the Copac Tool could fit into procedures for at-the-shelf inspection and what information would be useful. Generate use cases.
5. Potential Discards			
Set your own threshold number of duplicate copies held elsewhere and be able to also set a sliding scale.		<u>Criteria 1 & 2</u> Sliding threshold re duplication of particular value to Criteria 1, but could also be useful for Criteria 2 (e.g. to identify other institutions with especially strong/weak collections matching the batch) this could form part of identification of potential CCM	Seek feedback from WRC institutions on how the Copac Tool could fit into stock editing procedures, and what information would be useful.

Areas to look at	Comment	WRC Notes	Assessment
		partners for exchanging unwanted items in a given subject area.	
6. Output data			
csv output: consider fields output in light of potential uses and whether could be a batch output mode too. "MARC" format for import to LMS: required fields for "MARC" export need to be agreed.	Agreed that full MARC data is not required therefore use of Copac database is appropriate.	Export options to allow 'automated' tying of results back to local items using local control/circulation numbers. Would not anticipate including all but basic bibliographic data in downloads (e.g. author/title for quick accuracy checks) e.g. full MARC download not needed.	Seek feedback from WRC institutions on how the Copac Tool results would ideally be used to add information to local LMS records, and what format and information would be useful. Consider potential for using csv output for analysis via spread sheets

Appendix 2 – Phase 3 Testing Plan

Testing actions for Phase 3
Criterion 1 – Can I release this book?
a. Submit batch file of control numbers and download results (i.e. number of holdings per record) as MARC file (Use Case 1).
▪ Does the file contain the expected data fields?
▪ Can this data be imported into the LMS?
▪ Do the data fields provide the information necessary to inform stock management decisions?
▪ What further functionality would be useful?
b. Submit same batch file of control numbers and download results as comma-delimited file (if available).
▪ Does the file contain the expected data fields?
▪ Can this data be imported into Excel or Access?
▪ Do the data fields provide the information necessary to inform stock management decisions?
▪ What further functionality would be useful?
c. Submit same batch file of control numbers but limit results to specific libraries (White Rose ± British Library) for consortia stock management. Output could be either MARC or comma-delimited.
▪ Does the data provide a basis for stock management decisions within a consortia environment?
▪ Would additional data fields or functionality be helpful?
d. Test changes to the online interface.   To include:
▪ Data presentation in initial list of hits against the database (sort order, fields present, etc.)
▪ Data visualisation tools (esp. export of holding libraries as csv file)
▪ Can the search for a single item be embedded in other applications?
▪ Can we do any testing which simulates use cases 3 and 4?
e. Test functionality limiting results to specific libraries in the online interface (much as per above but with a single record number or ISBN).   To include:
▪ Are the data visualisation tools useful in this context?
▪ What further functionality would be useful?
▪ Consider applicability to use cases 3 and 4?
Criterion 2 - How does my collection compare in strength to other UK Libraries?
a. Submit large batch file of ISBNs (representing ca. 4,000 records from a particular subject area or collection) and evaluate results. Hopefully, this will include both numbers held by each library and number of holdings for each title.
▪ Are the results clear?
▪ Do they give useful information about the relationship between the collection submitted and national holdings?

	▪ Is the visualisation helpful (esp. for the holdings per title graph)?
	▪ How do the 2 graphs help to understand the collection?
	▪ What further questions arise from this analysis and how might one find answers?
	▪ What further functionality would be useful?
b.	Submit subject search and evaluate results. Hopefully, this will include both numbers held by each library and number of holdings for each title.
	▪ Which aspects of the visualisation are helpful in this instance?
	▪ Do the results give useful information about the national holdings in this area?
	▪ Does it provide any useful information about the relationship between local holdings and national holdings?
	▪ What further questions arise from this analysis and how might one find answers?
	▪ What further functionality would be useful?