

Understanding Collections Overlap

An investigation into White Rose Libraries
collections using Collection Management Tools

Final Report, July 2017

Summary

This report sets out the results of a Jisc-funded investigation by the White Rose Libraries (York, Leeds and Sheffield) to explore and validate the ways which the Jisc Copac Collection Management tool (CCM) and the SCS/OCLC GreenGlass tool attempt to match and de-duplicate bibliographic records; and how those results compare with manually checked results. The impetus for this work came from exercises carried out in 2016 using the GreenGlass tool that reported back a degree of overlap between collections that was much lower than anticipated. Jisc agreed to support a 'deep-dive' into the data on the basis that the results would be of broad interest to the library community. It was also clear that this work would usefully help refine collection management requirements, both for the ongoing development of the tools themselves, and for the emerging National Bibliographic Knowledgebase (NBK), the data from which is designed to provide a foundation for collection management activities in future.

Authors:

White Rose Library staff (see acknowledgements section)

Understanding collections overlap: an investigation into White Rose Libraries collections using the SCS GreenGlass and COPAC Collaboration Collection Management Tool

Final Report, July 2017

Understanding collections overlap: an investigation into White Rose Libraries collections using the SCS GreenGlass and COPAC Collaboration Collection Management Tool	1
Final Report, July 2017	1
Background	2
Understanding record matching in GreenGlass	3
Understanding record matching in the Copac CCM Tool	5
Summary of Data checking undertaken	5
Test 1	5
Test 2	6
Test 3	7
Test 4	8
Overview of outcomes	8
Recommendations	9
Recommendation 1: Share a version of this report more widely	9
Recommendation 2: Develop workflow guidance and best practice to help libraries export data to external catalogues	9
Recommendation 3: External catalogues should indicate the completeness and currency of contributing library holdings	9
Recommendation 4: Develop guidance to help libraries understand the impact that metadata quality has on matching records	9
Recommendation 5: Investigate ways to help libraries improve the quality of their metadata in catalogue records	10
Recommendation 6: Develop advice and guidance for libraries using collection analysis tools	10
Recommendation 7: Develop advice and guidance for libraries embarking on collaborative collection management initiatives	10
Recommendation 8: Contribute to the future development of collection analysis tools	10
Acknowledgements	11
Appendices	12

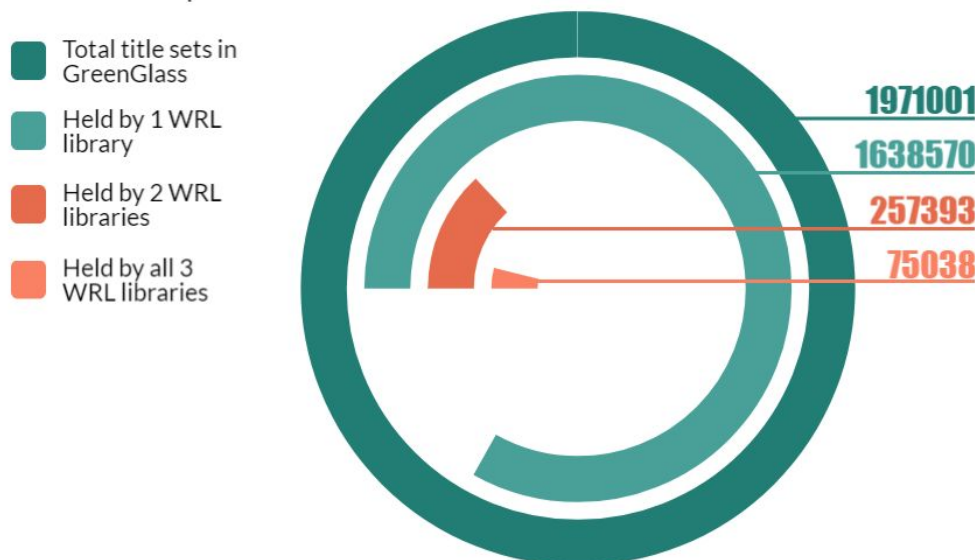
Background

In early 2016 the White Rose Libraries (at the universities of Leeds Sheffield and York) began work with GreenGlass to carry out an analysis of their collections in order to explore collaborative collection management between the 3 libraries. GreenGlass is a collection analysis tool developed by Sustainable Collection Services (SCS), who are now part of OCLC.

Catalogue records from the 3 White Rose libraries were loaded into GreenGlass in the summer of 2016. SCS analysed this metadata in terms of a range of factors such as circulation history; publication date; holdings across certain pre-agreed groupings of peer libraries. Classification was also 'normalized' across the three libraries' differing schemes by a SCS methodology which awarded a DDC number to each work. Results were received in Autumn 2016. The collection overlap reported by GreenGlass between the 3 White Rose Libraries (WRL) was considerably lower than expected:

- 83% of the titles in the 3 libraries' combined collections were identified as uniquely held by only 1 library (1,638,570 of 1,971,001 titles held)
- 75.5% of University of Leeds collections identified as unique within WRL
- 64.6% of University of Sheffield collections identified as unique within WRL
- 58.9% of University of York collections identified as unique within WRL

WRL shared print collection - title sets



We (the White Rose Libraries) therefore began conversations with SCS to try to gain a better understanding of how SCS/GreenGlass had uniquely identified records and otherwise processed the data supplied, in an attempt to verify whether the less than anticipated degree of overlap was correct. We also identified some in-depth checking work that we wished to carry out independently of SCS, comparing GreenGlass matching with the Copac Collection Management (CCM) Tool, as well as doing some manual checks on overlap; Jisc agreed to fund this work. We believed this work would be helpful to the UK library community as a whole, as well as to the White Rose Libraries as

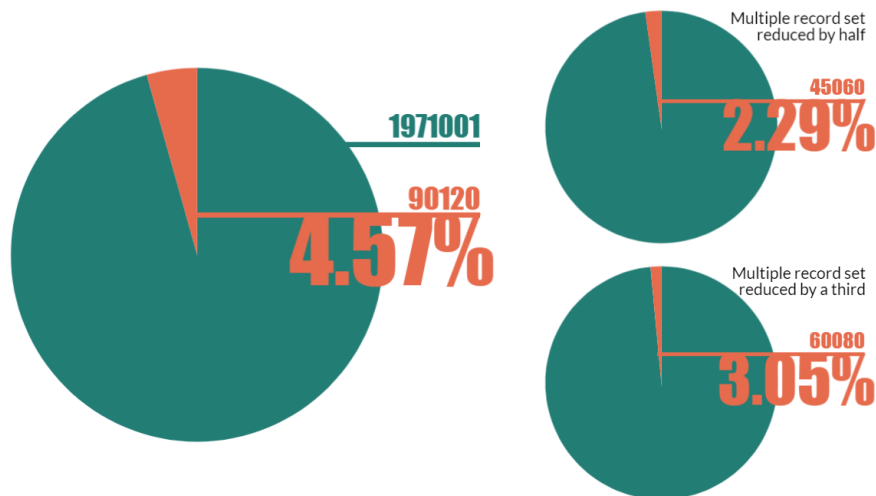
our study would explore the way in which both the CCM tool and GreenGlass operate their matching, and so give a better understanding of their usefulness in determining collection overlap. Benefit would also be gained from an understanding of the implications for use of GreenGlass across the UK Library community, some of which have already been identified by WRL, and an understanding of the CCM tool (of which there has been little analysis to date). With the move to the National Bibliographic KnowledgeBase (NBK), it is critically important to understand the requirement for collections analysis moving forward and our work should give an insight into necessary requirements and potential development pathways for collection analysis tools such as GreenGlass and the CCM Tool.

WRL will continue in their use of both GreenGlass and the CCM Tool beyond the lifespan of this project, further exploring the feasibility of collaborative collection management. WRL will continue to share their work with both the Jisc and the wider community.

Understanding record matching in GreenGlass

It is worth stating early on in this report that our data checking exercises have shown that GreenGlass has actually given us a reasonably accurate picture of the overlap between the WRL collections. It is also worth noting that the overlap between WRL is similar to that of US libraries who have already used the GreenGlass tool. We did identify that our overlap was slightly underreported, and established that the reasons for that underreporting were largely due to incomplete and inconsistent metadata in the WRL catalogue records.

GreenGlass uses the OCLC record number (the OCN) as the match point; this is present in catalogue records that have had their holdings set against WorldCat, and if not present it will be assigned programmatically by SCS during the data ingest. During the initial stages of our checking we discovered that different OCNs had been assigned to what we considered to be the same title held at different libraries. Ruth Fischer (SCS) found 90,120 titles (OCLC Work IDs) that contained multiple records (OCNs) for books published in the same year, a strong indicator that this was a set of records containing undetected duplicates. However, 90,120 of 1,971,001 is only 4.57% of the total number of titles; even if we were to assume this set could be reduced by between one third and one half, that does not greatly decrease our percentage uniqueness of titles held.



SCS have been very willing to work with us to consider whether alternative methods of matching might eliminate the underreporting of overlap. At Ruth Fischer's request, we checked in detail a small number of records from the set of 90,120 titles (OCLC Work IDs) that contained multiple records (OCNs) for books published in the same year. In many of these cases, White Rose Libraries have their WorldCat holding for the same title set on different records. Since SCS defines duplicates as those records that share an OCN (rather than a Work ID/title), these circumstances cause GreenGlass to understate overlap.

We checked these records to determine whether we believed they should be understood as duplicates within WorldCat. SCS have said they will incorporate our feedback as they continue to dig into the finer aspects of record matching.

[Summary of WRL testing](#) and [Detailed spreadsheet with results of WRL record checking](#).

This was a very useful exercise, as we began to gain real insight into how metadata quality, and historic cataloguing practices, can affect matching. We have begun to record that in this typology document: [Issues affecting accuracy of matching](#).

This work with SCS shows that using the OCLC Work ID (instead of the OCN currently used) would not be a more reliable way to identify duplicates. It succeeds in pulling together some records that are missed via OCN matching, but it also matches some that should not be considered duplicates. SCS is still investigating but think that the GLIMIR Content ID (which are described in this 2012 article: <http://journal.code4lib.org/articles/6812>) may be the answer. They will let us know what they learn from their investigations.

SCS have also observed that although a great deal of work is underway at OCLC to strengthen the WorldCat matching and de-duplication routines, they will remain complicated and the results imperfect.

Understanding record matching in the Copac CCM Tool

We consulted Shirley Cousins at Copac in order to gain a better understanding of record matching in the Copac database and, by extension, the CCM Tool. Shirley provided us with a helpful [document](#) which included the following summary:

“There is an initial match process that identifies potential duplicates. Matching records then go through a more detailed supplementary match process used to confirm or reject the initial match.

If the match between records is confirmed the records are merged to form a consolidated record. This creates a new record using data from the largest of the original records, also taking additional fields from the other matched records where appropriate eg. spelling variations in a title will be retained for indexing only, whilst additional subject terms will be included for both indexing and display. The consolidated record also includes holdings details for all the matched records. In addition, within the consolidation we retain each of the original records so that a consolidated record can be expanded to view all the records as originally supplied.

If a potential record match is rejected the new incoming record is added to Copac as a single, unconsolidated, record.”

As we had with GreenGlass, we discovered some examples of titles held by all three WRL that did not appear on just the one entry in COPAC and we sent these onto Shirley. She replied as follows:

“A quick check of a few of your ISBNs suggest that some match on ISBN but fail on other match elements - but some do look as though they should match. We can’t tell now why any particular match has failed - there may be a number of reasons for this to do with the state of the data at the time the records were added. We have an automated process we run from time to time that looks for additional duplicates that we’ve not identified in the initial load process, so we’ll re-start this and do some more matching. This set of ISBNs looks like it gives us some useful examples to work with, to consider how we might update the match process to pick up some additional duplicates.”

We look forward to receiving an update from Shirley about this in due course.

Shirley also explained that in the CCM Tool it was possible to request deduplication by ISBN, a good way of pulling records together that might otherwise have failed a Copac match for some reason. There are three levels of multi-field deduplication available on the CCM Tool: Level 1 uses Date, Title, Pagination, Edition, Author, Publisher, Level 2: uses Date, Title, Author, Publisher; Level 3: uses Title, Author.

The data checking exercises we conducted using the CCM Tool supported the level of overlap reported by GreenGlass.

Summary of Data checking undertaken

Test 1

We initially analysed the results generated by the overlap analyses facilitated by GreenGlass and compared them

against the CCM Tool in relation to 2 areas held by the White Rose Libraries (WRL), namely:

- Physics works with ISBNs
- Physics works without ISBNs

To do this, we produced a report of records identified by GreenGlass as being unique to one WRL within the WRL group. We then input these records into the CCM Tool, to see how many the CCM Tool identified as unique to one WRL, as well as how many were held by 2 or more WRL. As a further check, we reviewed the report of records from GreenGlass in Excel and manually calculated the number of records identified as unique to one WRL, as well as how many were held by 2 or more WRL.

The Physics results indicated:

- The presence or absence of an ISBN in the record has minimal impact (1-2%) on matching accuracy
- Manual Excel checking closely reflects the GreenGlass results, showing only a 2-4% difference from the GreenGlass totals
- The CCM Tool results differ by 11-12% difference from the GreenGlass totals. (We investigated this, and found that the discrepancy was due to records for items in York's External Store had not been exported to Copac)
- Overall therefore, the GreenGlass matching seems to differ by only small percentages from other methods tested
- However, small percentages translate into large numbers of books. Both WRL and SCS are therefore keen to understand more about the factors which inhibit matching (which prompted the 'Checking GreenGlass undetected duplicates' exercise)

Detail of the testing and our conclusions are available in the document [Overview of results for analysis of Physics \(Dewey 530\)](#).

We then moved on to doing a similar checking exercise with a different subject area, Art History, that is not so textbook-heavy (and York knew there would be few, if any, Art History books in their External Store!). We decided on Art History rather than French Literature, which had been the intention in our original proposal, because we realised the presence of diacritics was likely to have had significant impact on matching. The Art results were broadly similar to Physics.

Detail of the testing and our conclusions are available in the document [Overview of results for analysis of Art \(Dewey 700 - 710\)](#).

Test 2

Each library looked in detail at those items held by one other WRL (overlap = 2) or by two WRL (overlap = 3) from a range of subject areas, namely:

- Maths works with ISBNs
- Education works with ISBNs

- Chemistry works with ISBNs
- Physics works with ISBNs
- French Literature works with ISBNs
- Psychology works with ISBNs
- Linguistics works with ISBNs

Duplicate ISBNs were deleted and the number of records remaining for each subject area was noted.

The GreenGlass list was imported into the CCM Tool to identify Copac holdings for the listed ISBNs. The resulting report was then exported into Excel where the number of records held by the “home” WRL +1 (and +2) was identified through filtering the list.

The number of items reported by the CCM Tool was compared against that from GreenGlass, and results for each WRL compiled giving details for each subject area.

Detail of the testing and our conclusions are available in the document [Instructions for headline figures comparison \(GG/CCM\) 17/03/17](#). The results indicated that the CCM Tool appeared to report fewer overlapping titles than GreenGlass.

Test 3

We tested the difference in totals between the number of records entered into the CCM Tool (from an original GreenGlass sourced list), and the number of results which are produced as a result. For example, 100 record numbers may have been imported to the CCM tool, but results were produced for only 80. We wanted to understand which of the original records were not showing in the CCM Tool results and why that was, as well as examining the records which differed from GreenGlass in the manual spreadsheet.

There had been some discrepancies between the number of records imported into the CCM Tool (using either ISBN lists or lists of bibliographic record system numbers) and the number of results returned. During our initial testing of titles which GreenGlass had recorded as having no overlap across the WRL, the number of results exported from the CCM Tool were generally lower than the number imported.

For example for the Physics subject area (Dewey 530):

- 1139 records with ISBNs entered into the CCM Tool - 982 records exported from the CCM Tool
- 883 records without ISBNs entered into the CCM Tool - 581 records exported from the CCM Tool

We compared the lists of records imported into the CCM Tool with the lists of results exported, and tried to identify which records were missing. A sample revealed the missing records were in fact not currently in the Copac database (they were not included in the regular publishing job from York’s LMS to Copac). We were confident that this satisfactorily accounted for the difference in records.

Details of the testing are available in the document [Physics Testing Results](#).

A second example was tested for Art (Dewey 700-710).

We had imported a list of 8817 records with ISBNs into the CCM Tool (this was a list of records which GreenGlass recorded as being unique to the holding library), and 8769 records were exported. Upon investigation the disparity was caused by duplication of titles in the import record. There are instances where a particular library has multiple

bibliographic records for the same title (for example York Minster and University of York share a catalogue, but retain separate bibliographic and holdings records).

We were again satisfied that we understood the reasons for the discrepancy.

Details of the testing and our conclusions are available in the document [Art Testing Results](#).

Test 4

In the same way that we looked at WRL=1 for each library we wanted to explore what results WRL=3 in GreenGlass would produce. Subsequent testing by all three WRL showed that testing for an identical Dewey range with search criteria (WRL=3) did not produce an identical results for each of the WRL, as our initial assumption had been.

We discovered a number of different reasons for these anomalies:

- Internal duplication (more than one catalogue record for the same title in a library's system)
- Cataloguing differences, particularly for multi-volume sets
- Presence or absence of ISBNs
- Differences in Dewey numbers

Details of the testing and our conclusions are available in the document [GreenGlass lists WR = 3](#).

Overview of outcomes

To summarise very briefly, there appear to be 3 main categories of factors affecting the matching of WRL records, which apply when analysis is carried out with either of the tools looked at here. We believe these are worth flagging to Jisc as things to consider in the NBK project.

1. Data preparedness and exports: the profiles for regular exports to external catalogues (Copac, WorldCat, etc) need to be understood and re-assessed when a new project is undertaken, rather than simply copied across. It is very easy for libraries to lose sight/understanding of what decisions were taken about which records to export, and why. If exports for the NBK simply replicate what was set up for Copac they may be incomplete, as we have found. The amount of uncatalogued material should also be understood and highlighted by each institution so that a national picture of 'hidden' collections can be uncovered, as well as the proportion of the collection classified in Dewey vs other schema, including local ones.
2. Metadata quality: not only does current cataloguing practice vary between UK libraries, each library will also have records reflecting a variety of different legacy approaches to cataloguing. In addition, UK libraries obtain their downloaded records from a number of different sources. Data migration between library systems, as well as records ingested due to organisational mergers, are also likely to have had an impact. The typology document we referred to earlier, [Issues affecting accuracy of matching](#), gives examples of metadata variations which can prevent matching e.g. ISBNs for different editions within the same record, presence of qualifiers (pbk.), differences in name entries, in titles, abbreviations, publication details, size, series, print/e on the same record, different practices for cataloguing multi volume sets.

3. Matching algorithms: it is extremely unlikely that any automated approach to matching could ever be 100% accurate, but we have found it is worth taking time to investigate how matching works in different tools, and to encourage the providers of those tools to experiment with different algorithms. Sometimes we felt that the matching algorithms might be too precise, for example the presence/absence of diacritics and symbols.

Recommendations

Recommendation 1: Share a version of this report more widely

White Rose Libraries have welcomed the opportunity to carry out this detailed 'data digging', and believe that our experiences, observations and conclusions could not only assist other libraries who want to use these tools, but also benefit aspects of the National Bibliographic Knowledgebase work.

Recommendation 2: Develop workflow guidance and best practice to help libraries export data to external catalogues

One of the main causes of anomalies uncovered by our checking work was the differences in the export files the same library had sent to WorldCat, Copac and GreenGlass. It seems obvious to state that data sets should be consistent and complete, but we uncovered a lack of shared common knowledge between metadata, collection analysis and systems staff about the detail of what was exported to external catalogues, and the impact that would have on collection analysis work. For example, a decision had been taken at one site some years previously to exclude stock in an off site store from the Copac export because it only contained journals; when books were added to that store at a later date the export profile was not updated. The unforeseen consequence was that that stock was therefore not available to be analysed by the CCM tool. Developing clear and accessible best practice guidelines would help other libraries avoid this sort of pitfall, and would be especially timely considering the data exports to the NBK that are currently underway.

Recommendation 3: External catalogues should indicate the completeness and currency of contributing library holdings

Following on from the above, NBK and other external catalogues should make it clear what contributing libraries have included/excluded, and make it easy to see how current the holdings for each contributing library are. Libraries embarking on collaborative print initiatives, especially those making retention and disposal decisions in that context, need to be confident that they understand the data they are analysing about other libraries' collections.

Recommendation 4: Develop guidance to help libraries understand the impact that metadata quality has on matching records

Libraries would benefit from a clear understanding of which fields are more/less influential in record matching. As already mentioned, we have started to create a typology document: [Issues affecting accuracy of matching](#).

It would also be helpful to offer the option at record ingest for the library to specify fields/subfield to be stripped out/ignored in the matching process; in Alma, for examples, we can use rules to filter out unwanted metadata fields/subfields, but this may not be the case for all systems.

Recommendation 5: Investigate ways to help libraries improve the quality of their metadata in catalogue records

The variance and range in the quality of metadata in records is a nationwide problem in the UK, in comparison with the more centrally sourced records in the US. This will be a challenge to the NBK work in relation to matching records. One option could be to encourage or facilitate libraries to carry out some small scale 'data improvement' projects, to evaluate possible methods and then see whether this has improved matching. Another option would be to offer the facility for libraries to receive back from the NBK the "master" record that their record has matched with; this would allow libraries to choose if they want to import/merge/overlay the NBK master record (or key fields from it) into their own system. To ensure accurate matching, the record received would ideally contain the unique system number from the contributing library's repository.

Recommendation 6: Develop advice and guidance for libraries using collection analysis tools

We have established that both GreenGlass and CCM Tool work in very different ways and operate using different parameters. Our explorations have given us some understanding of how each tool works that we would be happy to share with other libraries. It would also be useful at the outset for any library using such a tool to have an understanding of how the matching is working, and this would be greatly facilitated if collection analysis tools were transparent about the way in which their matching operates. For example, what you intend to use the tool for might influence what holdings you decide to upload into GreenGlass: if a library wants to identify titles to withdraw, we would suggest uploading only the material you would be willing to dispose of; however, for more general collections analysis work a much broader range of material should be uploaded.

Another learning point for us about GreenGlass was that it does not do any internal deduplication or normalisation, leading to duplication of titles within the same library's results set. A library may have duplicate records for the same title, which may or may not be valid current practice, but when doing the analysis work it's important to understand the impact of this on some reports.

Recommendation 7: Develop advice and guidance for libraries embarking on collaborative collection management initiatives

In addition to sharing our experiences with others, WRL would also like to understand how library consortia in the US have progressed their collaborative collection management schemes using tools such as GreenGlass. We understand there are consortia using GreenGlass to manage such schemes, so it would be useful to understand what those schemes are setting out to achieve, and what kinds of working arrangements have been put in place. We would also like to gauge the degree to which the presumed dependency of US academic libraries in OCLC WorldCat as a single metadata source lends itself to the accuracy and effectiveness of GreenGlass as a collection analysis tool.

Recommendation 8: Contribute to the future development of collection analysis tools

WRL have several suggestions for ways in which tools like GreenGlass and CCM might be developed, which we would like to discuss with interested parties. For example:

- Options to select whether precise or more fuzzy matching is required. Libraries may be prepared to accept different levels of risk (of imprecise matching) depending on the nature of the work being carried out, and the significance of a particular collection.
- It may also be helpful to be given options to determine which fields to match on (in other words increased transparency in the way in which the tool works, and increased control then over how the matching is implemented).
- If a tool is to be used for collaborative collection management, then it is helpful to see which libraries are holding a copy of a particular item which is in several locations. At present in GreenGlass it is not possible to know which the other holding libraries are.
- In order for a library to be able to use a collection management tool to perform stock editing work on its own collections, it needs to be able to work within the tool using its own classification scheme directly, without translation mapping of the local scheme. Currently this is not possible in GreenGlass (but this development has been promised).
- A GreenGlass remediation report showing when a library has the only holding set against an OCN, when there are other OCNs with the same publication date; this would indicate a discrepancy in the record which needs correcting.

White Rose Libraries would be keen to continue discussions with Jisc and work with them and other relevant partners to take these recommendations forward.

Acknowledgements

With thanks to Jisc for the funding which enabled our detailed data checking and the production of this report, and to Shirley Cousins (Jisc) and Ruth Fischer (SCS) for answering our questions about Copac/CCM Tool and GreenGlass respectively.

The following people should be credited for this report and the associated work.

University of Leeds

Ian Jennings, Jane Edwards, Jane Saunders

University of Sheffield

Amanda Doherty, Andy Bussey, Chris Ashton, Fran Abbs, Gary Ward, Tracey Clarke

University of York

Liz Waller, Matt Wigzell, Ruth Elder, Sarah Thompson, Sue Elphinstone

White Rose Libraries Executive

Kate Petherbridge, Tom Grady

All the documents linked to from this report are available in an [Appendices folder](#).

Appendices

Appendix 1_ Understanding collections overlap: an investigation

Appendix 2_ Summary of testing - records with the same OCLC work ID

Appendix 3_ LEEDS Record checking -work IDs with multiple OCNs that share a pub year

Appendix 3_ SHEFFIELD Record checking -work IDs with multiple OCNs that share a pub year

Appendix 3_ YORK Record checking - work IDs with multiple OCNs that share a pub year

Appendix 4_ Issues affecting accuracy of matching

Appendix 5_ Copac record match summary 0217

Appendix 6_ Overview of results for analysis of Physics Dewey 530 Test 1

Appendix 7_ Overview of results for analysis of Art Dewey 700 - 710 Test 1

Appendix 8_ Instructions for headline figures comparison GGCCM 170317 Test 2

Appendix 9_ Physics ISBN Testing Test 3

Appendix 10_ Art ISBN Testing Test 3

Appendix 11_ Greenglass lists WR 3 Test 4

Appendix 12_ Issues affecting accuracy of matching

Understanding collections overlap: an investigation into White Rose Libraries collections using the SCS Greenglass and COPAC Collaboration Collection Management Tool

Lead contact

Jane Saunders

University of Leeds

Leeds

LS2 9JT

Project partners

Liz Waller

University of York

York

YO10 9JT

Tracey Clarke

University of Sheffield

Sheffield

S10 3TN

Project start date: 13th March 2017

Project end date: 31st July 2017

Total funding requested: £25,000

Understanding collections overlap: an investigation into White Rose Libraries collections using the SCS Greenglass and COPAC Collaboration Collection Management Tool

Background

In early 2016 the White Rose Libraries (at the universities of Leeds Sheffield and York) began work with GreenGlass to carry out an analysis of their collections in order to explore collaborative collection management between the 3 libraries. Records were loaded from the 3 libraries into GreenGlass in the summer of 2016, and results received in Autumn 2016. The collection overlap identified by GreenGlass between the 3 White Rose libraries is considerably lower than expected:

- 83% of the titles in the 3 libraries' combined collections are identified as uniquely held by only 1 library (1,638,570 of 1,971,001 titles held)
- 75.5% of University of Leeds collections identified as unique within WRL
- 64.6% of University of Sheffield collections identified as unique within WRL
- 58.9% of University of York collections identified as unique within WRL

Conversations with Greenglass are ongoing to gain a better understanding of the data. However, the libraries feel that investigations into the results of GreenGlass matching will be helpful not just to the White Rose Libraries but to the community as a whole, and that it would be useful to compare these results against those achieved when the Copac Collection Management (CCM) Tool is used. Additionally it is proposed that a manual check on overlap is also carried out. The White Rose libraries also feel that the low level of overlap reported against the British Library holdings by the GreenGlass data (53% of holdings of the White Rose libraries are reported as not held by the BL) merits investigation.

White Rose Libraries (WRL) have already invested in SCS Greenglass as a tool for collection analysis in monetary terms (costs for data upload and analysis) and in kind, through the time used in analysis to date. WRL will continue in their use of both Greenglass and the CCM tool beyond the lifespan of this project and will share their work with both the Jisc and the wider community. WRL intend to work towards a shared print collection and collaborative retention policies.

The work undertaken by WRL will benefit the wider community in verifying the overlap between the holdings of the three libraries and with British Library holdings (and therefore potential for overlap in the UK). Benefit will also be gained from an understanding of the implications for use of Greenglass across the UK Library community, some of which have already been identified by WRL, and an understanding of the CCM tool (of which there has been little analysis to date). With the move to the National Bibliographic Knowledge Base, it is critically important to understand the requirement for collections analysis moving forward and the work proposed will give an insight into necessary requirements and potential development pathways for Greenglass and the CCM tool.

Caroline Brazier, Chief Librarian of The British Library, has offered to run our data against British Library holdings. We do not anticipate the BL requiring funding for this work, though there will be a cost to WRL in preparing the data. We will be formalising this arrangement with Caroline in due course.

Initial discussions were held with OCLC and SCS on the 28th February and both have offered their assistance with the project. Brief discussions have also taken place with Diana Massam and Shirley Cousins regarding the CCM tool. Further input is likely to be required from COPAC/CCM staff, probably equivalent to 3 days over the lifetime of the project. No further input is anticipated from Jisc, although it may be useful to have regular contact between project milestones.

Timeline

This is the initial timeline for the project. This will be discussed further at the kick off meeting on 16th March, after which a more detailed breakdown will be used.

	Meetings	Processing and overall analysis of GreenGlas s and CCM Tool results from all sites	Analysis of overlap results	Manual check of holdings	Analysis of results and preparatio n of report	Analysis with BL
Week 1 w/c 13th March						
Week 2						
Week 3						
Week 4						
Week 5						
Week 6						
Week 7						
Week 8						
Week 9						
Week 10 w/c 15th May						
Week 11						
Week 12						

Week 13						
Week 14						
Week 15						
Week 16						
Week 17						
Week 18						
Week 19						
Week 20 w/c 24th July						

Milestones and deliverables

Milestones will be marked by the three meetings: a kick off meeting with WRL will take place on Thursday 16th March at the University of York, a mid-point meeting will be held w/c 15th May (date to be arranged) and a closing meeting w/c 24th July. Representatives from Jisc are welcome to attend these meetings either by Skype or in person.

The WRL project group will meeting fortnightly via Skype.

The deliverable for this project will be a full report to JISC by 31st July 2017. This report will set out the level of overlap found by the 3 approaches (GreenGlass, CCM Tool and the manual check), and will include an analysis of the findings, including references to detailed analyses of sets of bibliographic records of items included, and the results of running holdings in the WRL datasets used against those of the BL. Any files used through the project will be retained, and can be made available to JISC for further examination if this is required.

Risks

Risk	Mitigation
Staffing not adequate to fulfill the project	Named staff have been identified by each institution with funding allocated to backfill
Lack of understanding of Greenglass and CCM tool	Close involvement of SCCS and CCM tool colleagues to inform investigations
Timescale inadequate to complete the work	WRL have undertaken preliminary work on collection analysis and this experience leads us to believe that the timescale is realistic

There are risks involved in not undertaking this project. With work underway to deliver the NBK this project can provide insights into the quality of data available, mitigating risk in the creation of the database. In addition the community currently relies on the CCM tool, and it is necessary moving forward to have some form of tool which to use in association the the NBK. The work undertaken will provide a basis for discussion on the fitness for purpose of Greenglass and the CCM tool and inform development of a future mechanism for collection management analysis.

Project dissemination

An appropriate version of the report could be made publically available and project staff would be available to attend conferences and meetings to discuss the work

Method

The method analyses the results generated by the overlap analyses facilitated by the CCM Tool and the GreenGlass software in relation to 4 areas held by the White Rose Libraries, namely:

- Physics works with ISBNs
- Physics works without ISBNs
- French Literature works with ISBNs
- French Literature works without ISBNs

For each of these areas a manual check on overlap will also be conducted.

The activity will be coordinated by York, with each of the White Rose library sites providing files from their own systems, and each of the libraries variously contributing to the costs of record analysis and the manual checking of overlaps.

Costs

Activity	Site of Activity	Person rate	Costs Sites of Activity
Processing and overall analysis of GreenGlass and CCM Tool results from all sites; coordination of in depth record analysis (20-25 days total)	York	£200	£5000
Uploading of files and liaison with York (5 days total)	Leeds / Sheffield	£200	£1000
Manual checking of holdings for overlaps for items without ISBNs (8 days total)	Leeds/Sheffield	£200	£1600

Analysis of bib records to examine variations in overlap results (40 days) (shared across sites depending on capacity)	Leeds / Sheffield / York	£200	£8100
Summary of findings / conclusions written (10 days)	Leeds / Sheffield / York	£200	£2000
Analysis / liaison with BL on overlap with BL records, plus reporting on results (10 days)	One site to lead	£200	£2000
Senior manager time (9 days)	Leeds / Sheffield / York	£370	£3300
Administration of costs (1 day)	Leeds	£200	£200
Meetings and travel (most meetings will use Google Hangouts, but some will require face to face work)	Leeds / Sheffield / York	£200	£1800
Total			£25,000

Jane Saunders, Leeds University Library
Liz Waller, University of York Library
Tracey Clarke, University of Sheffield Library

Summary of testing - records with the same OCLC work ID

[Record checking -work IDs with multiple OCNs that share a pub year](#)

The purpose was to check groups of records which have the same OCLC work ID, but have different Worldcat OCLC numbers, and indicate whether we believe they should be understood as duplicates. (GreenGlass uses Worldcat OCLC numbers rather than OCLC work ID for matching).

York's findings

York checked 10 groups of records. Eight of these ten contain records which we would want to be considered a match. Within these eight are two examples (see rows 11-13 and 35-38) where we would want two out of three records to match, but not the other. Rows 11-13: York's copy has a different publisher and ISBN to Leeds & Sheffield. Rows 35-38: Leeds has a different ISBN to Sheffield and York, and seems to be a different edition.

The non-matching within the other two groups (lines 2-10 and 17-20) can be explained by differences in cataloguing practice (e.g. multi-volume items which one institution has catalogued individually, whilst another has only 1 bib record). It is worth noting that multi-volume sets do therefore present a challenge.

During the checking we noted down some of the discrepancies between records which have not been assigned the same Worldcat OCLC numbers. We think it's possible that differences in the way the publisher, place of publication, edition statement etc has been recorded causes records to be assigned different numbers.

Sheffield's findings

At Sheffield we checked 20 records. There were 13/ 20 given a "yes" result to indicate they are the same.

Those grouped together by OCLC work ID which are not the same were for

- works with different formats (e.g. microform, eBook, hard copy) This makes me curious as to which field the format type is taken from?

- those with different publishers, imprints and editions (some have vague or ambiguous metadata and look the same but aren't) This highlights how a lot of “differences” are based on qualitative data.
- multivol. sets. (largely because Sheffield used to catalogue each one rather than have one bib. record for the whole set with items attached) . This causes problems in the title field as well as strange ISBN matching. How will this apply to bound volumes is a concern, could this lead to a false match with a “part” of a locally bound volume?

Those grouped together which are the same but did not originally match suggest the following :

- authorisable fields matter. If we don't all use the same form of name it seems to cause a non-match.
- whether there is a creator or title main entry for the same work main entry seems to make a difference. Much cataloguing work is open to interpretation so we may not all agree on which of these we should use.
- strange displays of diacritics in title field could lead to problems with matching. Are they being read as totally different words if the coding is out?
- series data matters. Does the presence of the now defunct 440 field instead of the 490 field prevent a match? Or is it an accompanying 830 that's required?

In general it suggests that local cataloguing practices and lack of standardised data is behind a lot of the reasons for any original mismatches

Leeds' findings

I think I'm still not very clear on the difference between a 'bib_oclc_nbr' and a 'worldcat_oclc_nbr'. Does the latter apply to a manifestation of a work, i.e. the same work published under two different imprints are different manifestations and therefore should have different worldcat OCLC numbers? Is the former just a unique identifier for a bibliographic record in the OCLC scheme of things?

In other respects, Leeds also checked the SCS/GreenGlass matching decisions for 10 OCLC work id's. It was concluded that the matching outcome at the work level was correct in 9 instances. Matching outcomes at the manifestation level were correct for 8 works.

The above tallies include the two print and microform examples in the sample assigned to Leeds. These are the same work, but different manifestations. Bringing the manifestations together (separately) under work ids 10022889:eng & 10025319:lat, but giving them different OCLC numbers (as per GG) and different WorldCat numbers, suggests their differences have also been correctly noted. Issues as to how these records are then organised in WorldCat perhaps start to crop up with

10022889:eng when the WorldCat holdings display indicates Leeds does not have the microform version when, in fact, we do.

There is certainly one work in the sample where it is evident the matching process at work level has failed:

Work id 10026710:eng The work under OCLC bib no 17607485 should not be under this work id. This perhaps demonstrates how the matching algorithm can be satisfied too soon or too easily leading to an incorrect outcome - same date; same author, but a different work.

Works in the sample where work id is correctly assigned but the manifestations could be considered different (therefore requiring different OCLC record numbers) are:

1002472:eng and 10031735:eng (unless, in the latter case, they are actually both the International edition).

I would suggest these findings may bring us back to the earlier question of, in our collections analysis, what degree of difference between works and manifestations we require these tools to observe and report. Presumably we would want for a monograph and microform version of exactly the same text to be regarded as different, but a (change of imprint) reprint where there are no differences in text, pagination, typesetting, etc to be regarded as the same?

oclc_work_id	Same? Y/N	Discrepancies	pub_year	edition	bib_oclc_nbr	worldcat_oclc_nbr	inst_name	worldcat_evidenc e_type	bib_title	bib_author	publisher	isbn
10022889:eng	N (On grounds of one is print, the other microform)	300 c notes octavo size; has 600 entry; chosen topical subject headings differ from microform record. Curious as to how microforms are in this sample as they should have been excluded from the dataset sent to SCS. Certain commas absent from 245; 300 pagination descriptive styling differs, no note of size/format; no 600 entry.	1699		15313183	559094403	Univ of Leeds		The Christian ministry of the Church of England vindicated and distinguished from the antichristian ministry of the Quakers [microform] : containing a brief reply to a false and foolish libel stiled A letter to the clergy of the diocese of Norfolk and Suffolk, &c., by a nameless author ... wherein his folly is detected, his lies confuted ... / by a member of the Church of England, Francis Bugg.	Bugg, Francis, 1640-1724?	London : Printed for the author, and are to be sold by J. Robinson ... and H. Rhodes ..., 1699.	
10022889:eng	N (On grounds of one is print, the other microform)	Arguable as to whether the microform version is a distinct manifestation and therefore should have a different (reprint?) pub. date.	1699		25792944	933089985	Univ of Leeds		The Christian ministry of the Church of England vindicated and distinguished from the antichristian ministry of the Quakers : containing a brief reply to a false and foolish libel, stiled, A letter to the clergy of the diocese of Norfolk and Suffolk, &c., by a nameless author ... wherein his folly is detected, his lies confuted ... / by a member of the Church of England, Francis Bugg.	Bugg, Francis, 1640-1724?.	London : Printed for the author, and are to be sold by J. Robinson ... and H. Rhodes ..., 1699.	
1002472:eng	N	Whilst textually these might well be the same, they have different imprints. However, worldcat.org would appear to lump the holdings together into the one entry under a slightly different imprint from that for this Methuen publication.	2002		48979714	48979714	Univ of Sheffield	1.00	Speaking Shakespeare / Patsy Rodenburg.	Rodenburg, Patsy, 1953-	London : Bloomsbury Methuen Drama, c2002. [New York] ; Houndmills : Palgrave Macmillan, 2002.	9780413762702
1002472:eng	N		2002		50479899	48979714	Univ of York		Speaking Shakespeare / Patsy Rodenburg.	Rodenburg, Patsy, 1953-		9780312294205
10025319:lat	N (On grounds of one is print, the other microform)	Abbreviated title, with differences in spaces and comma punctuation, in 245; no 300	1685		13672732	Can't find this entry in worldcat.org	Univ of Leeds		Defensio fidei Nicaenae [microform] : ex scriptis, quae extant, Catholicorum doctorum, qui intra tria prima ecclesia Christianae secula floruerunt : in qua obiter quoque Constantinopolitana confessio, de Spiritu Sancto, antiquiorum testimoniis adstruitur / authore Georgio Bullo ...	Bull, George, 1634-1710.	Oxonii : E Theatro Sheldoniano, 1685.	
10025319:lat	N (On grounds of one is print, the other microform)	245 has subtitle; 260 [b has s.n. rather than place name; 300 is recorded in some detail	1685		907590583	Can't find this entry in worldcat.org	Univ of York		Defensio fidei Nicaenae exscriptis quae extant catholicorum doctorum, etc.	Bull, George, 1634-1710	Oxonii, : e theatro Sheldoniano, 1685.	
10025480:ita	Y	Incorrect placing of 245 apostrophe; 245 no subtitle differentiation; 260 has attempt at place of publication; 300 lack detail compared to the above record	1966		831313705	3339875	Univ of Leeds		Scritti sull'ebraismo in memoria di Guido Bedarida.		Firenze : [Bet-ha-ari], 1966.	
10025480:ita	Y		1966		941039688	3339875	Univ of Leeds		Scritti sull'ebraismo : in memoria di Guido Bedarida.		Firenze : [s.n.], 1966.	
1002567:eng	Y	Two ISBNs are hbk & pbk editions. Only other difference is Sheffield record has copyright date in addition to publication date. Note also, Leeds has a copy of this but under 10-digit ISBN and with pub date 2000. Is this not observed by overlap at the other two libraries?	2004		44786158	473917570	Univ of York		Green screen : environmentalism and Hollywood cinema / David Ingram.	Ingram, David, 1959-	Exeter : University of Exeter Press, 2004. Exeter : University of Exeter Press, 2004, c2000.	9780859896085
1002567:eng	Y		2004		56460828	473917570	Univ of Sheffield		Green screen : environmentalism and Hollywood cinema / David Ingram.	Ingram, David, 1959-		9780859896092
10026710:eng	Y	245 has statement of responsibility; 260 c Actual publication date (1967) in square parentheses; no 300 field	1966		3460303	740509697	Univ of Leeds		Marcellus Laroon / Robert Raines.	Raines, Robert.	London : Paul Mellon Foundation for British Art : Routledge & Kegan Paul, 1966 [i.e.1967]	
10026710:eng	Y	1966 recorded as publication date.	1966		154150105	740509697	Univ of Sheffield		Marcellus Laroon / Robert Raines.	Raines, Robert.	London : Paul Mellon Foundation for British Art : Routledge & Kegan Paul, 1966.	
10026710:eng	Y		1967		3460303	740509697	Univ of Leeds		Marcellus Laroon / by Robert Raines.	Raines, Robert.	London : The Paul Mellon Foundation for British Art [in association with] Routledge & K. Paul, 1967.	
10026710:eng	Y	245 has no statement of responsibility; 260: 1967 recorded as definite publication date; 300 present which goes into some detail	1967		3460303		Univ of York		Marcellus Laroon.	Raines, Robert	London, etc., Routledge & Kegan Paul, etc., 1967.	
10026710:eng	N	This is an exhibition catalogue, much less extensive (44 pages vs. 219) compared to the preceding 4 works. Matching process appears to have been 'fooled' by fact the two works share the author and publication date. Nevertheless, the exhibition catalogue should not be under this work ID.	1967		17607485	314616909 (possibly)	Univ of Leeds		Marcellus Laroon : an exhibition of paintings and drawings arranged by the Paul Mellon Foundation for British Art.	Laroon, Marcellus, 1679-1772.	[London] : Paul Mellon Foundation for British Art, [1967]	

oclc_work_id	Same? Y/N	Discrepancies	pub_year	edition	bib_oclc_nbr	worldcat_oclc_nbr	inst_name	worldcat_evidenc e_type	bib_title	bib_author	publisher	isbn
1003067:eng	Y	For SCS to give Leeds holding a pub_year of 2005 is incorrect when its record states 2004. Sheffield would appear to have 2005 as an incorrect pub date?	2005		56685114	656786270	Univ of Leeds		Making sense of children's drawings / John Willats.	Willats, John.	Mahwah, NJ : L. Erlbaum Associates, 2004.	9780805845389
1003067:eng	Y		2005		61872680	656786270	Univ of Sheffield		Making sense of children's drawings / John Willats.	Willats, John.	Mahwah, N.J. : Lawrence Erlbaum, 2005.	9780805845372
1003100:eng	Y	Author's date of birth; different level of detail in the 245 statement of responsibility.	1977		2646229		Univ of Sheffield		Attitudes and opinions / (by) Stuart Oskamp, in collaboration with ... (others).	Oskamp, Stuart, 1930-	Englewood Cliffs ; London (etc.) : Prentice-Hall, 1977.	9780130503930
1003100:eng	Y	York record has series statement.	1977		252314342		Univ of York		Attitudes and opinions; [by] Stuart Oskamp, in collaboration with Catherine Cameron, Mark W. Lipsey, Burton Mindick [and] Theodore Weissbach.	Oskamp, Stuart	Englewood Cliffs, Prentice-Hall, 1977.	
10031735:eng	Y	The only difference of note is the use of 'international' in York's series statement.	2001	4th ed.	45058829		Univ of Sheffield		Applied hydrogeology / C.W. Fetter.	Fetter, C. W. (Charles Willard), 1941 ;	Upper Saddle River, N.J. : Prentice Hall, 2001.	9780130882394
10031735:eng	Y		2001	4th International	223111013		Univ of York		Applied hydrogeology / C. W. Fetter.	Fetter, C. W. (Charles Willard), 1941 ;	Upper Saddle River, NJ : Pearson Education, c2001.	9780131226876
1003261:eng	Y	[New ed.] statement absent from Leeds record at time of investigating.	1954		1228446		Univ of Leeds		Commerce of the prairies / edited by Max L. Moorhead.	Gregg, Josiah, 1806-1850.	Norman : University of Oklahoma Press, [1954]	
1003261:eng	Y		1954	[New ed.]	813232307		Univ of Sheffield		Commerce of the Prairies / edited by Max L. Moorhead.	Gregg, Josiah, 1806-1850.	Norman : University of Oklahoma Press, 1954.	

oclc_work_id	Same? Yes/No	Discrepancies	pub_year	edition	bib_oclc_nbr	worldcat_oclc_nbr	inst_name	worldcat_evidence_type	bib_title	bib_author	publisher	isbn
10004648:eng	Yes	700 10 \$a McDermott, Richard A. \$q (Richard Arnold) 700 10 \$a Snyder, William.\$d 1956-	2002			48083908	Univ of York	3.00	Cultivating communities of practice : a guide to managing knowledge / Etienne Wenger, Richard McDermott, William M. Snyder.	Wenger, Etienne, 1952-	Boston : Harvard Business School Press, 2002.	9781578513307
10004648:eng	Yes	Matched with above on 48083908	2002			48083908	Univ of Leeds	3.00	Cultivating communities of practice : a guide to managing knowledge / Etienne Wenger, Rochard McDermott, and William Snyder.	Wenger, Etienne, 1952-	Boston, Mass. : Harvard Business School ; London : McGraw-Hill, 2002.	9781578513307
10004648:eng	Yes	700 10 \$a McDermott, Richard. 700 10 \$a Snyder, William	2002		847459117	847459117	Univ of Sheffield	1.00	Cultivating communities of practice : a guide to managing knowledge / Etienne Wenger, Richard McDermott, William Snyder.	Wenger, Etienne, 1952-	Boston, Mass. : Harvard Business School ; London : McGraw-Hill, 2002.	9781578513307
10006415:eng	No	Microfilm 100 1 \$a Weldon, Anthony, \$c Sir, \$d 1583?-1648. 700 1 \$a Howell, James, \$d 1594?-1666. Not a microfilm , it's an eBook Two forms of author Weldon, Anthony for the eBook and Weldon , Anthony, Sir, d. 1649? For the microform	1659		13523605	13523605	Univ of Leeds	1.01	A perfect description of the people and countrey of Scotland [microform]	Weldon, Anthony, Sir, d. 1649?	London : Printed for Rich. Lownds, 1659.	
10006415:eng	No		1659		99825309	222439096	Univ of Leeds	4.01	A perfect description of the people and country of Scotland [microform]	Weldon, Anthony, Sir, d. 1649?	London : printed for J.S., 1659.	
1000660:fre	Yes		1963			271358	Univ of York	4.01	LAUTREAMONT AND SADE	Blanchot, Maurice		
1000660:fre	Yes	Title in caps and in English . \$a Blanchot, Maurice, \$d 1907-600 10 \$a Ducasse, Isidore Lucien, \$d 1846-1870. \$t Chants de Maldoror. 600 10 \$a Sade, \$c marquis de, \$d 1740-1814. Diacritics ? 100 10 \$a Blanchot, Maurice, \$d 1907-600 10 \$a Ducasse, Isidore Lucien, \$d 1846-1870. 600 10 \$a Sade, \$c marquis de, \$d 1740-1814 700 11 \$a Lautréamont, \$c comte de, \$d 1846-1870 \$t Chants de Maldoror.	1963			3171436	Univ of Leeds	4.00	Lautr��amont et Sade, avec le texte int��gral des Chants de Maldoror.	Blanchot, Maurice.	[Paris] : ��ditions de Minuit, [1963]	
1000660:fre	Yes	100 10 Blanchot, Maurice. 600 14 \$a Sade, \$c marquis de, \$d 1740-1814. 700 02 (2nd Indicator) Lautr��amont, \$c comte de,\$d 1846-1870. \$t Chants de Maldoror.	1963		277228451	277228451	Univ of Sheffield	1.00	Lautre��x0081_amont et Sade : avec le texte int��x0081_gral des Chants de Maldoror [par le comte de Lautrel��x0081_amont] / Maurice Blanchot.	Blanchot, Maurice.	Paris : El��x0081_dition s de Minuit, 1963.	
10007058:eng	Yes	100 form of name differs from below (see name format placing of the word Sir) Budge, E. A. Wallis, Sir (Ernest Alfred Wallis), 1857-1934.	100			4614099	Univ of Leeds	2.00	Facsimiles of Egyptian hieratic papyri in the British Museum : with descriptions, translations, etc. / by E.A. Wallis Budge	Budge, E. A. Wallis, Sir (Ernest Alfred Wallis), 1857-1934.	London : British Museum, 1910	
10007058:eng	Yes	100 form of name differs from above (see name format placing of the word Sir) Budge, E. A. Wallis (Ernest, Alfred Wallis), Sir, 1857-1934	1910			931252521	Univ of York	4.00	Facsimiles of Egyptian hieratic papyri in the British Museum with descriptions, translations, etc. By E.A. Wallis Budge, ...	Budge, E. A. Wallis (Ernest, Alfred Wallis), Sir, 1857-1934	London, sold at the British Museum; and at Longmans & Co., Bernard Quaritch, Asher & Co.; and Henry Frowde, Oxford University Press, London, 1910.	
10007604:eng	Yes	Form of names differs - Maurice, of Sully, Bishop of Paris, ca. 1120-1196 and Robson, Charles Alan.	1952			3385756	Univ of Leeds	4.00	Maurice of Sully and the medieval vernacular homily : with the text of Maurice's French homilies, from a Sens Cathedral Chapter ms / by C.A. Robson.	Maurice, of Sully, Bishop of Paris, ca. 1120-1196.	Oxford : Basil Blackwell, 1952.	
10007604:eng	Yes	Maurice, of Sully, Bishop of Paris, approximately 1120-1196 and Robson, Charles Robson but by C. A. Robson. in 245	1952		926822134	926822134	Univ of Sheffield	1.00	Maurice of Sully and the medieval vernacular homily / by C.A.Robson ; with the text of Maurice's French homilies from a Sens Cathedral ms ; [ed.] by C. A. Robson.	Maurice, of Sully, Bishop of Paris, approximately 1120-1196.	Oxford : Blackwell, 1952.	

10009856:eng	No	Differing formats- e.g. microform and book	1631		99843798	55183705	Univ of Leeds	4.01	The English dictionarie or, An interpreter of hard English words [microform] : enabling as well ladies and gentlewomen, young schollers, clerkes, merchants; as also strangers of any nation, to the vnderstanding of the more difficult authors already printed in our language, and the more speedy attaining of an elegant perfection of the English tongue, both in reading, speaking, and writing. The third edition, reuised and enlarged. By H.C. Gent.	Cockeram, Henry, fl. 1650.	London : Printed by Thomas Harper, for Thomas Weauer, and are to be sold at his shop, at the great north dore of Pauls Church, 1631.	
10009856:eng	No		1631			228714171	Univ of Leeds	4.01	The English dictionarie or, An interpreter of hard English words : enabling as well ladies and gentlewomen, young schollers, clerkes, merchants; as also strangers of any nation, to the vnderstanding of the more difficult authors already printed in our language, and the more speedy attaining of an elegant perfection of the English tongue, both in reading, speaking, and writing. The third edition, reuised and enlarged / By H.C. Gent.	Cockeram, Henry, fl. 1650.	London : printed by Thomas Harper, for Thomas Weauer, and are to be sold at his shop, at the great north dore of Pauls Church, 1631.	
10009856:eng	No	Microform	1632		99853836	55192799	Univ of Leeds	4.01	The English dictionarie. Or, an interpreter of hard English words [microform] : enabling as well ladies and gentlewomen, young schollers, clerkes, merchants; as also strangers of any nation, to the vnderstanding of the more difficult authors already printed in our language, and the more speedy attaining of an elegant perfection of the English tongue, both in reading, speaking, and writing. The fourth edition, reuised and enlarged. By H.C. Gent.	Cockeram, Henry, fl. 1650.	London : Printed by Thomas Harper, for Thomas Weauer, and are to be sold at his shop, at the great North dore of Pauls Church, 1632.	
10009856:eng	No	Book	1632			606539653	Univ of Leeds	4.01	The English dictionarie. Or, an interpreter of hard English words : enabling as well ladies and gentlewomen, young schollers, clerkes, merchants; as also strangers of any nation, to the vnderstanding of the more difficult authors already printed in our language, and the more speedy attaining of an elegant perfection of the English tongue, both in reading, speaking, and writing. The fourth edition, reuised and enlarged / By H.C. Gent.	Cockeram, Henry, fl. 1650.	London : Printed by Thomas Harper, for Thomas Weauer, and are to be sold at his shop, at the great North dore of Pauls Church, 1632.	
10009856:eng	No	Some slightly different wording in the titles - eg "merchants" and "mercants [sic]"	1651	The tenth edition, revised and enlarged.	19719176	19719176	Univ of Leeds	1.01	The English dictionarie, or, An interpreter of hard English words [microform] : enabling as well ladies and gentlewomen, young scholars, clerks, mercants [sic] as also strangers of any nation to the understanding of the more difficult authors already printed in our language and the more speedy attaining of an elegant perfection of the English tongue both in reading, speaking and writing / by H.C. ...	Cockeram, Henry, fl. 1650.	London : Printed by W. Bentley, and are to be sold by Andrew Crook in S. Paul's Church-yard, at the sign of the Green Dragon, 1651.	
10009856:eng	No		1651	The tenth edition, revised and enlarged.	19719411	19719411	Univ of Leeds	1.01	The English dictionarie, or, An interpreter of hard English words [microform] : enabling as well ladies and gentlewomen, young scholars, clerks, merchants as also strangers of any nation to the understanding of the more difficult authors already printed in our language and the more speedie attaining of an elegant perfection of the English tongue both in reading, speaking and writing / by H. C. ...	Cockeram, Henry, fl. 1650.	London : Printed by W. Bentley, in part of recompense from A. Miller ..., 1651.	
10010580:eng	No	Hard copy and Form of name Norton, Thomas	1570			4402883	Univ of York	4.00	A bull graunted by the Pope to Doctor Harding & other, by reconcilement and assoyling of English Papistes, to vndermyne faith and allegiance to the Quene. With a true declaration of the intention and frutes thereof, and a warning of perils therby imminent, not to be neglected.	Norton, Thomas	Imprinted at London : By Iohn Daye, [1570].	
10010580:eng	No	Ebook and microform items. (RLUK record is an EBBO eBook record) Form of name Norton, Thomas, 1532-1584.	1570		99856916	55155884	Univ of Leeds	4.01	A bull graunted by the Pope to Doctor Harding & other [microform] : by reconcilement and assoyling of English Papistes, to vndermyne faith and allegiance to the Quene. With a true declaration of the intention and frutes thereof, and a warning of perils therby imminent, not to be neglected.	Norton, Thomas, 1532-1584.	Imprinted at London : By Iohn Daye dwelling ouer Aldersgate, [1570]	

1001229935:eng	Yes		2012		747534691	747534691	Univ of Leeds	1.00	The night of broken glass : eyewitness accounts of Kristallnacht / edited by Uta Gerhardt and Thomas Karlauf ; translated by Robert Simmons and Nick Somers.	Nie mehr zurÃ¼ck in dieses Land. English	Cambridge, United Kingdom ; Malden, Massachusetts : Polity Press, 2012.	9780745650845
1001229935:eng	Yes	Extra authors on this record not on above	2012	English ed.	802177360	802177360	Univ of Sheffield	1.00	The night of broken glass : eyewitness accounts of Kristallnacht / edited by Uta Gerhardt and Thomas Karlauf ; translated by Robet Simmons and Nick Somers.		Cambridge : Polity Press, c2012.	9780745650845
10013070:eng	Yes	Can't get into record from Leeds catalogue, but there doesn't seem to be an ISBN	1976			3414621	Univ of Leeds	2.00	Voluntary social service manpower resources / [by Adrian Webb, Lesley Day, Douglas Weller].	Webb, Adrian, 1943-	London : Personal Social Services Council, [1976]	
10013070:eng	Yes	Sheffield record has publication date in brackets	1976		877503596	877503596	Univ of Sheffield	1.00	Voluntary social service manpower resources / (by) Adrian Webb, Lesley Day, Douglas Weller.	Webb, Adrian, 1943-	London (2 Torrington Place, WC1E 7HN) : Personal Social Services Council, (1976).	9780905250021
10015518:eng	Yes	Different forms of name	3386077	Univ of Leeds	4.00	Memoir of William Tanner / compiled chiefly from autobiographic al memoranda ; edited by John Ford.	Tanner, William, 1815-1866.	London : F. Bowyer Kitto, 1868.				
10015518:eng	Yes	Different forms of name	931127995	Univ of York	4.11	The memoir of William Tanner : compiled chiefly from autobiographic al memoranda / edited by John Ford.	Tanner, William, 1815-1866.	London : Bowyer Kitto ; York : William Sessions, 1868.				
10015518:eng	Yes	Different forms of name. Added place of publication and publisher	931127995	Univ of York	4.00	Memoir of William Tanner, compiled chiefly from autobiographic al memoranda; ed. by John Ford. Preface by Sarah W. Tanner.	Tanner, William	London, F. Bowyer Kitto; York, William sessions; 1868.				
10016090:eng	Yes	George Lawless in 100 field	15631246	Univ of Sheffield	1.00	Augustine of Hippo and his monastic rule / George Lawless.	Lawless, George.	Oxford : Clarendon Press, 1987.	9780198266877			
10016090:eng	Yes	As above George Lawless in 100 field	15631246	Univ of Leeds	3.00	Augustine of Hippo and his monastic rule / George Lawless.	Lawless, George.	Oxford : Clarendon Press, 1987.	9780198266877			
10016090:eng	Yes	Has Augustine of Hippo as author, George Lawless in 700 field. Also form of publisher name is slightly different	26302037	Univ of York	3.00	Augustine of Hippo and his monastic rule; [ed. and tr. by] George Lawless, OSA.	Augustine, Saint, Bishop of Hippo	Oxford, Clarendon P., 1987.	9780198267416			

1001647:eng	Yes	Sheffield uses the defunct 440 for series and not 490 / 830 and so looks like it's not part of any series?	30974080	Univ of Leeds	3.00	The evolution of the sailing navy, 1509-1815 / Richard Harding.	Harding, Richard, 1953-	Basingstoke : Macmillan, 1995.	9780312124076
1001647:eng	Yes	Uses 49010 and 830 fields for series Also has c before the date.	60225814	Univ of Sheffield	1.00	The evolution of the sailing navy, 1509-1815 / Richard Harding.	Harding, Richard, 1953-	Basingstoke : Macmillan, c1995.	9780333596043
10016539:eng	No	Different publisher and place of publication. Different form of name	15661435	Univ of Leeds	3.00	More die of heartbreak : a novel / Saul Bellow.	Bellow, Saul, 1915-2005.	London : Alison Press, 1987.	9780436039621
10016539:eng	No	Different publisher and place of publication. No author in 245 field. Different form of name	56548010	Univ of Sheffield	1.00	More die of heartbreak.	Bellow, Saul, 1915-	London : Secker and Warburg, 1987.	9780436039621
10017029:eng	Yes	Egmont on record only as additional creator. Publisher details different	15464935	Univ of Leeds	4.00	Things as they are.		London : Printed for S. Hooper, and A. Morely, 1758.	
10017029:eng	Yes	Publisher details described differently but it appears to be the same?	931192502	Univ of York	4.00	Things as they are.	Egmont, John Perceval, Earl of, 1711-1770	London, for S. Hooper & A. Morley, G. Woodfall, & J. Staples, 1758.	
1001712:eng	No	2005 ed (updated and enlarged from 1998 ed.)	62217294	Univ of Leeds	3.00	Pakistan : a modern history / Ian Talbot.	Talbot, Ian.	London : C. Hurst, c2005.	9781850653851
1001712:eng	No	2005 ed (updated and enlarged from 1998 ed.)	926795033	Univ of Sheffield	1.00	Pakistan : a modern history / Ian Talbot.	Talbot, Ian.	London : Hurst, 2005.	9781850653851
1001712:eng	No	2009 ed.	38739043	Univ of Leeds	2.00	Pakistan : a modern history / Ian Talbot.	Talbot, Ian.	London : Hurst, 2009.	9781850659891
1001712:eng	No	2009 ed.	806096917	Univ of York	3.00	Pakistan : a modern history / Ian Talbot.	Talbot, Ian.	London : Hurst, 2009.	9781850659891

10017387:eng	No	Published Texas. Different ISBNs	3414759	Univ of Sheffield	1.00	Arms and the wizard : Lloyd George and the Ministry of Munitions, 1915-1916 / R. J. Q. Adams.	Adams, R. J. Q., 1943-	College Station : Texas A&M University Press, c1978.	9780890960455
10017387:eng	No	Published by Cassell, London. Different ISBNs	4047914	Univ of Leeds	3.00	Arms and the wizard : Lloyd George and the Ministry of Munitions, 1915-1916 / R. J. Q. Adams.	Adams, R. J. Q., 1943-	London : Cassell, 1978.	9780304299164
10019282:eng	Yes	Different form of name for Nathaniel Bacon. Publisher looks different, though I guess it's the same one.	3339639	Univ of Leeds	4.00	The official papers of Sir Nathaniel Bacon of Stiffkey, Norfolk, as justice of the peace, 1580-1620 / selected and edited for the Royal Historical Society from original papers formerly in the collection of the Marquess Townshend, by H.W. Saunders.	Bacon, Nathaniel, Sir, 1547-1622.	London : Offices of the Society, 1915.	
10019282:eng	Yes	Different form of name for Nathaniel Bacon. Publisher looks different, though I guess it's the same one.	57402941	Univ of York	4.00	The official papers of Sir Nathaniel Bacon, of Stiffkey, Norfolk, as Justice of the Peace, 1580-1620. / Selected and edited for the Royal Historical Society from original papers formerly in the collection of the Marquess Townshend, by H. W. Saunders.	Bacon, Nathaniel, 1547-1622.	London : Royal Historical Society, 1915.	
10020441:eng	Yes	Strange... Can't see why original mismatch as they are both 1990 paperback eds. but there's a ebook attached to this too. Is there some eBook metadata distinguishing it from the Sheffield copy?	17806205	Univ of Leeds	3.00	Dewey / J.E. Tiles.	Tiles, J. E.	London : Routledge, 1990.	9780415053105
10020441:eng	Yes	See above, bizarrely Sheffield also has 1988 copy with 17806205 (above for Leeds) ???	39525322	Univ of Sheffield	1.00	Dewey / J. E. Tiles.	Tiles, J. E.	London : Routledge, 1990.	9780415053105

10043:eng	No	See above	13790835	Univ of Leeds	The morning ran Payne, Henry Ne London : Printed for Thomas Dring ..., 1673.
-----------	----	-----------	----------	---------------	--

[illegible]

oclc_work_id	Same? Y/N	Discrepancies	pub_year	edition	worldcat_oclc_ nbr	inst_name	bib_title	bib_author	publisher	isbn
10035003:ger	N	Leeds version catalogued on 1 bib record - Sheffield's have individual records (one for each volume)	1914		271095245	Univ of Leeds	Ausgewählte Werke / Martin Luther ; unter Mitwirkung von Hermann Barge ... [et al.], herausgegeben von Hans Heinrich Borchardt.	Luther, Martin, 1483-1546.	München : Georg Müller, 1914-1925.	
10035003:ger	N	Separate record for each volume	1914		926801949	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.2, Reformatorische und politische Schriften: die grossen Reformationsschriften von 1520.	Luther, Martin, 1483-1546.	München : Müller, 1914.	
10035003:ger	N	Separate record for each volume	1922		270810242	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.3, Reformatorische und politische Schriften: aus den Tagen des Wormser Reichstags.	Luther, Martin, 1483-1546.	München : Müller, 1922.	
10035003:ger	N	Separate record for each volume	1922		926801984	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.1., Reformatorische und politische Schriften: der Ablassstreit und die Leipziger Disputation.	Luther, Martin, 1483-1546.	München : Müller, 1922.	
10035003:ger	N	Separate record for each volume	1923		270810246	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.6, Schriften zur Neuorganisation der Gesellschaft; Der grosse Katechismus.	Luther, Martin, 1483-1546.	München : Müller, 1923.	
10035003:ger	N	Separate record for each volume	1923		926801609	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.4, Reformatorische und politische Schriften: der Kampf gegen Schwarm- und Rottengeister.	Luther, Martin, 1483-1546.	München : Müller, 1923.	
10035003:ger	N	Separate record for each volume	1923		926802068	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.5, Vom unfreien Willen; Schriften zur Neuorganisation der Kirche.	Luther, Martin, 1483-1546.	München : Müller, 1923.	
10035003:ger	N	Separate record for each volume	1925		271095245	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.7, Predigten; Vermischte Schriften; Dichtungen.	Luther, Martin, 1483-1546.	München : Müller, 1925.	
10035003:ger	N	Separate record for each volume	1925		926849586	Univ of Sheffield	Ausgewählte Werke / herausgegeben von H.H. Borchardt. Bd.8, Tischreden.	Luther, Martin, 1483-1546.	München : Müller, 1925.	
10035302:eng	N	Different Publisher: Bloomington : Indiana University Press, c1978. ISBN 0253340594	1978		3415326	Univ of York	Nicholas I, emperor and autocrat of all the Russias / W. Bruce Lincoln.	Lincoln, W. Bruce.	Bloomington : Indiana University Press, c1978.	9780253340597
10035302:eng	Y	London : Allen Lane, 1978. (No note in Sheffield record.) Different ISBN to York ISBN 0713908378; ISBN 9780713908374	1978		4184095	Univ of Sheffield	Nicholas I : Emperor and autocrat of all the Russias / W. Bruce Lincoln.	Lincoln, W. Bruce.	London : Allen Lane, 1978.	9780713908374
10035302:eng	Y	London : Allen Lane, 1978. (Note in Leeds record: Also published: Bloomington : Indiana University Press, 1978.) Different ISBN to York. ISBN0713908378	1978		4184095	Univ of Leeds	Nicholas I, emperor and autocrat of all the Russias / W. Bruce Lincoln.	Lincoln, W. Bruce.	London : Allen Lane, 1978.	9780713908374
1003640004:eng	Y	Print only	2012		747713165	Univ of Sheffield	Loverly : the life and times of My fair lady / Dominic McHugh.	McHugh, Dominic.	New York : Oxford University Press, c2012.	9780199827305
1003640004:eng	Y	Print only	2012		747713165	Univ of Leeds	Loverly : the life and times of My fair lady / Dominic McHugh.	McHugh, Dominic.	New York : Oxford University Press, c2012.	9780199827305
1003640004:eng	Y	Print and E copy held at York - on same bib record.	2012		868924032	Univ of York	Loverly : the life and times of My fair lady / Dominic McHugh.	McHugh, Dominic.	New York : Oxford University Press, c2012.	9780199827312
10036705:fre	N	Different volumes within same series. Series title is same on each record, but sub'title is for individual volumes.	1928		277536096	Univ of Sheffield	Anthologie des poètes français des origines à nos jours. t.1, De la Cantilène d'Eulalie à Pierre Ronsard.	Mazade, Fernand, 1863-1939.	Paris : Librairie de France, [1928]	
10036705:fre	N		1928		277536123	Univ of Sheffield	Anthologie des poètes français des origines à nos jours. t.2, De Joachim Du Bellay à Pierre Corneille.	Mazade, Fernand, 1863-1939.	Paris : Librairie de France, [1928]	
10036705:fre	N		1928		277536127	Univ of Sheffield	Anthologie des poètes français des origines à nos jours. t.3, De Scudéry à mme Desbordes-Valmore.	Mazade, Fernand, 1863-1939.	Paris : Librairie de France, [1928]	
10036705:fre	N		1928		277536149	Univ of Sheffield	Anthologie des poètes français des origines à nos jours. t.4, De Lamartine à Verlaine.	Mazade, Fernand, 1863-1939.	Paris : Librairie de France, [1928]	
1003777985:eng	Y	Only difference I can identify is double entry in Leeds record of Series information: SeriesOxford applied linguistics Oxford applied linguistics.	2011		747816093	Univ of Leeds	Understanding English as a lingua franca / Barbara Seidlhofer.	Seidlhofer, Barbara.	Oxford : Oxford University Press, 2011.	9780194375009
1003777985:eng	Y	Series: Oxford applied linguistics	2011		759841770	Univ of Sheffield	Understanding English as a lingua franca / Barbara Seidlhofer.	Seidlhofer, Barbara.	Oxford : Oxford University Press, 2011	9780194375009
1003777985:eng	Y	Series: Oxford applied linguistics.	2011		759841770	Univ of York	Understanding English as a lingua franca / Barbara Seidlhofer.	Seidlhofer, Barbara.	Oxford : Oxford University Press, 2011.	9780194375009
10038844:eng	Y	Assume that they are the same, but there are significant differences in the records: in title entry (Leeds inc author), in publisher location (London/Cambridge). Looking at each individual record, Sheffield has no ISBN number or physical details where as Leeds does.	1966		3460806	Univ of Sheffield	Causes of the slow rate of economic growth of the United Kingdom : an inaugural lecture.	Kaldor, Nicholas, 1908-1986.	London : Cambridge University Press, 1966.	
10038844:eng	Y		1966		643109206	Univ of Leeds	Causes of the slow rate of economic growth of the United Kingdom : an inaugural lecture / by Nicholas Kaldor.	Kaldor, Nicholas, 1908-1986.	Cambridge : Cambridge University Press, 1966.	9780521054621
1004072344:eng	Y	Yes, same. Leeds & Sheffield have country designator after place of publication. York doesnt	2011	2nd ed.	213113442	Univ of York	What is nursing? : exploring theory and practice / Carol Hall, Dawn Ritchie.	Hall, Carol, RGN.	Exeter : Learning Matters, 2011.	9780857254450
1004072344:eng	Y		2011	2nd ed.	747917289	Univ of Sheffield	What is nursing? : exploring theory and practice / Carol Hall, Dawn Ritchie.	Hall, Carol, RGN.	Exeter [England] : Learning Matters, 2011.	9780857254450
1004072344:eng	Y		2011	2nd ed.	747917289	Univ of Leeds	What is nursing? : exploring theory and practice / Carol Hall, Dawn Ritchie.	Hall, Carol, RGN.	Exeter [England] : Learning Matters, 2011.	9780857254450

oclc_work_id	Same? Y/N	Discrepancies	pub_year	edition	worldcat_oclc_ nbr	inst_name	bib_title	bib_author	publisher	isbn
10040795:eng	Y	Title: one has comma, one doesn't. Publisher: \$\$b is different	1924		15513722	Univ of Leeds	Unemployment, 1920-1923.	International Labour Office.	Geneva : [Printed by A. Kundig], 1924.	
10040795:eng	Y		1924		277459778	Univ of Sheffield	Unemployment 1920-1923.	International Labour Office.	Geneva : [s.n.], 1924.	
1004094:eng		Sheffield has catalogued some of the volumes individually	1965		4729711	Univ of Sheffield	The collected works of Walter Bagehot / edited by Norman St. John Stevas. vol.1, The literary essays (in two volumes) / with an introduction by Sir William Haley.	Bagehot, Walter, 1826-1877.	London : The Economist, 1965.	
1004094:eng	Y		1965		4729711	Univ of York	The collected works of Walter Bagehot; ed. by Norman St. John Stevas.	Bagehot, Walter, 1826-1877	London, The Economist, 1965-86.	
1004094:eng	Y		1965		4729711	Univ of Leeds	The collected works of Walter Bagehot / edited by Norman St John Stevas.	Bagehot, Walter, 1826-1877.	London : Economist, 1965.	
1004094:eng		Volume catalogued individually	1965		270724663	Univ of Sheffield	The collected works of Walter Bagehot / edited by Norman St. John Stevas. Vol.2, The literary essays (in two volumes).	Bagehot, Walter, 1826-1877.	London : The Economist, 1965.	
1004180:eng	N	6th Internation ed (does not state as revised.) ISBN - 908911, 0814	2013	6th ed., Internatio nal ed.	799139378	Univ of Leeds	Using multivariate statistics / Barbara G. Tabachnick, Linda S. Fidell.	Tabachnick, Barbara G., 1936-	Boston, Mass. : London : Pearson, c2013.	9780205890811
1004180:eng	Y	Print and e-version on same record. 6 Revised ed. E book ISBN - 4546, print - 1317	2013	Internatio nal ed of 6th revised ed.	855890781	Univ of York	Using Multivariate Statistics / Tabachnick, Barbara G.		Harlow : Pearson Education, 2013.	9781292021317
1004180:eng	Y	6th edition. New international edition - 1317.	2014	Sixth edition. Internatio nal	62766132	Univ of Sheffield	Using multivariate statistics / Barbara G. Tabachnick, Linda S. Fidell.	Tabachnick, Barbara G., 1936- author.	Harlow, Essex : Pearson, [2014]	9781292021317
1004180:eng	Y	Why are there two entries for Sheffield and on different numbers? According to Star cat, there is only one bib entry for this pub date. ISBN - 1317	2014	Sixth edition. Internatio nal	855890781	Univ of Sheffield	Using multivariate statistics / Barbara G. Tabachnick, Linda S. Fidell.	Tabachnick, Barbara G., 1936- author.	Harlow, Essex : Pearson, [2014]	9781292021317
				Note: slight level of concern about more recent publication dates not matching					Note: It appears that Place of publication has been included as part of matching process. This may differ, or could include country as well as place in some records. It may be present in some records or catalogued as [s.n.] in others. There are too many variables for place to be included in matching process.	

Differences WRL have encountered in testing that might have affected matching in OCLC, GreenGlass or Copac



Typology of metadata issues

ISBNs

- ISBNs for different editions within same record

- Common practice to add e-book ISBNs to print records (& vice versa) could be problematic for matching

- Presence of qualifiers (pbk) / (hbk) following ISBN

- 13- / 10- digit ISBNs

Differences in name entries

- e.g. Oskamp, Stuart, 1930- (Sheffield) & Oskamp, Stuart (York)

Differences in titles

- Multi volume works catalogued by series title or individual vols

- Punctuation e.g. “Unemployment, 1920-1923” & “Unemployment 1920-1923” not matched

- Titles lacking statement of responsibility e.g. Marcellus Laroon / by Robert Raines (Leeds) & Marcellus Laroon (York)

- Additional names added to statement of responsibility e.g. translated by ...,

- Titles in capitals (York)

Presence of diacritics, symbols & abbreviations

- York used [] in titles

Differences in Publication places, publishers & dates

- Use of more than one place of publication

- Country designator included in one record but not another

- [s.n.] used in one record, when place recorded in other

- Different UK /US publishers for same title

- Publication date discrepancies

Differences in recorded size

- 21cm / 24cm - **why use size as a match criteria?**

- Pagination - **do differences in page numbers result in poor matching? Do we have examples?**

Series

- More than one series title recorded in a record

- Series titles recorded in 440 tag or 830 tag

Other issues

Print and ‘e’ recorded on same record

Copac Record Matching: Summary

February 2017

The following provides a brief summary of the record match procedure used to create the Copac database. There is an initial match process that identifies potential duplicates. Matching records then go through a more detailed supplementary match process used to confirm or reject the initial match.

If the match between records is confirmed the records are merged to form a consolidated record. This creates a new record using data from the largest of the original records, also taking additional fields from the other matched records where appropriate eg. spelling variations in a title will be retained for indexing only, whilst additional subject terms will be included for both indexing and display. The consolidated record also includes holdings details for all the matched records. In addition, within the consolidation we retain each of the original records so that a consolidated record can be expanded to view all the records as originally supplied.

If a potential record match is rejected the new incoming record is added to Copac as a single, unconsolidated, record.

1. Identifying potential matches

Incoming records go through an initial match process that checks for potential duplicates by matching new records against those records already in the database using the Title and Date indexes. Record pairs that are identified as potential matches on the basis of their title then go into the more detailed Supplementary Match process that is used to confirm or reject this initial potential match.

2. Supplementary match procedure

The Supplementary Match process confirms or rejects the output of the initial potential duplicates match process. Which route the records take through this more detailed match procedure depends on an initial standard number match and/or the nature of the material described in the record.

Record pairs containing Standard Number (SN) elements generally go through a Quick Match. This speeds the matching process and also avoids having to match on some of the less consistent elements such as publisher. Other records go through the Detailed match process.

2.1 Standard Number match 1

If any of the following Standard Number (SN) match 1 checks are true the record pair goes through the Quick match process, otherwise the record pair goes through the Full match process.

- Two periodical records with at least one matching ISSN
- OR All ISBN's match
- OR All ISMN's match
- OR All ESTC numbers match.

2.2 Quick Match process

For records that have passed SN match 1, the Quick match checks the record pair for matching title and edition. If this match succeeds the duplicate record pair is confirmed and the records become part of a consolidation. If the match fails the incoming record is added to the database as a single, unconsolidated, record.

2.3 Detailed Match process

Records that fail SN match 1 go through the Detailed Match process. If the record pair fails any test the match process ceases. If the record pair passes all the match tests they are confirmed as duplicates and the records become part of a consolidation.

2.3.1 Standard Number match 2

A second Standard Number (SN) check, SN match 2, is used to identify the route the record pair takes though the Full match tests. This time the SN match only requires one SN in common between the records. The ISSN match is a failsafe to pick up any records that have managed to get through with an ISSN that are not identified as periodicals. Unlikely but not impossible.

- Do the records have ISBN, ISMN, ISSN or ESTC numbers in common?

This check assigns a flag to the record pair that either lets it through just the basic match tests, or forces it through the additional tests required where there is no SN in common between the records.

2.3.2 Match procedure

- If both records have an ISSN or ISMN or ISBN or ESTC number, do they have one in common?
- Are there more than 4 ISBN's? If so do they match?
Merging records for single volumes of sets with multi-volume records is potentially problematic. But we want to be able to match records where one has, say, ISBN's for paperback and hardback whilst the other has only the paperback ISBN.
- Do the dates match?
This is *not* used where both records in a pair are periodicals.
Uses 008, 260, 264.
- If both records are *periodicals* do the hierarchical places match?
Uses 752. This is primarily for matching some newspaper records.
- Do the titles match?
This checks 245 title as well as volume for multi-part works. It uses a fuzzy match allowing for minor variation, but preserving single letter 'words'. A smaller subset of subfields are used for matching periodicals. Includes checks for more complex title, edition and statement of responsibility details, including title truncation, in pre-1800 works and older records.
- Do the editions match?
Matches word and number variants.
- Do the series volumes match?
Uses the 440 if present, or 490.
- Do the authors match?
Corporate author stopwords are removed and there is a fuzzy match process that allows for some minor variation. The match uses 1XX and 7XX fields. If the usual author fields are not present it will check the 130, 730, 720, 245.

If the records failed SN match 2 then the following additional tests are used:

- Do the pages match?
Uses the 300. This is *not* used where both records in a pair are periodicals.
- Do the publisher names match?
Uses the 264, 260. Common stopwords are excluded and there is a partial match on publisher name and/or location depending on the information available.
- Do the map scales match?
Uses the 034.
- Do the music score types match?
Uses the 300, 240, 245. It checks for a range of score types eg. choral score.

Overview of results for analysis of Physics (Dewey 530)

Headline Information

From the results (listed in Table 1), there is only 1-2% points difference in the accuracy matching between those records reported in GG with ISBNs and without ISBNs and those reported within a specific testing method (i.e. within either Local Testing or CCM Tool testing.)

It had been anticipated that there would be a much greater difference between with/without ISBN, with the assumption that the testing would be much more accurate against the records containing ISBNs; this has not been the case.

Local testing (manual Excel checking) closely reflects the GG results, showing only a 2 - 4% difference from the GG totals.

Matching on the same title/author/edition in GG appears to have failed on the occasions when there are discrepancies in the records of individual libraries in the formatting of the author and/or the publication details. It also appears to fail to match through irregular use of punctuation (for example the use of square brackets or non standard abbreviations.) *

The CCM Tool results have between 11 - 12% difference from the GG totals.

(York records imported into the CCM Tool which did not produce a result through the CCM Tools were identified and examined. On testing - all of these items are held at the York External Store.)

Matching in GG of Non ISBN stock has failed for the same reasons as listed above when compared with matching through the CCM and manual checking.*

Matching (or nonmatching) of ISBN stock has been investigated in document [ISBN Testing \(Individual Titles\) York](#).

From these results it could be concluded that - dependent on acceptable level of risk to the library/collaboration - the GG results are reliable enough to move forward on (with an understanding of their matching criteria.)

Report of work

Methodology in brief:

- To produce a report of records identified by GreenGlass (GG) as unique to one WRL within the WRL group.
 - To input the records identified by GG into the CCM tool, and calculate the number of records identified as unique to one WRL within the WRL group. (Also figures for records held by 2 or more WRL.)
 - To review the report of records from GG in Excel and manually calculate the number of records identified as unique to one WRL within the WRL group. (Also figures for records held by 2 or more WRL.)
-

Review of overall results

GreenGlass results:

Physics (DDC 530) Unique to one WRL within the WRL group.

- Total with ISBN = 5357
- Total without ISBN = 5794

CCM tool results:

(Using GG report as original source of data for input file to tool)

- With ISBN - identified as unique to one WRL out of the WRL group:
 $4734/5320 = 89\%$
- With ISBN - identified as held by 2 or more WRL out of the WRL group:
 $586/5320 = 11\%$
- Without ISBN - identified as unique to one WRL out of the WRL group:
 $4677/5300 = 88\%$
- Without ISBN - identified as held by 2 or more WRL out of the WRL group:
 $623/5300 = 12\%$

Local Environment Checking:

- With ISBN - identified as unique to one WRL out of the WRL group:
 $5137/5357 = 96\%$
- With ISBN - identified as held by 2 or more WRL out of the WRL group:
 $220/5357 = 4\%$

- Without ISBN - identified as unique to one WRL out of the WRL group:
5704/5794 = 98%
- Without ISBN - identified as held by 2 or more WRL out of the WRL group:
90/5794 = 2%

Method of testing	Records with ISBN	Records without ISBN
Greenglass total	5357	5794
Local testing total	96% of GG total (5137 records) (-220 from GG total)	98% of GG total (5704 records) (-90 from GG total)
CCM Tool testing	89% of GG total (4734) (-623 from GG total)	88% of GG total (4677) (-1117 from GG total)

Table 1

During the testing of the above work York realised that data inputted into GG includes External Store Book stock data. However - on checking - the External Store Book stock data is not currently exported into Copac (originally York did not keep book stock in the External Store.) This impacts on the difference in Yorks data reports between GG and Copac - on occasions producing results through the CCM tool which are significantly lower in number than the GG totals. . This is particularly pronounced in the Science subject areas (most of the book stock in the store is Science related.)

Note: in future monthly exports this data will now be included by York.

Suggested next actions

(The complete process for checking Physics (DDC 530) data has been checked twice and appears to be accurate within the known limitations)

I would suggest that the entire process is now re-run for an arts/humanities based subject area (English language) to see if a similar pattern of results are produced.

(York holds minimal arts/humanities book stock in the External store.)

Comments

As well as needing to have a very clear understanding of how both OCLC and COPAC matching works, it is also equally important to have a very clear understanding of any discrepancies between GG data loads and what is surfaced in Copac. This means comprehending exactly which specific library location/collections are imported into the respective databases by each library in order to ensure like is being matched with like.

Overview of results for analysis of Art (Dewey 700)

(Updated 14.30: 18.05.17)

**Review of number of records identified as unique to one WRL within the WRL group
(Art - Dewey 700 - 710)**

Table 1

Method of testing	Records with ISBN	Records without ISBN
Greenglass total	8817 records	6952 records
Local testing total	8413 records (95% of GG figure for records with ISBNs) (-404 from GG total)	6656 records (96 % of GG figure for records without ISBN) (-296 from GG total)
CCM Tool testing	90 % of records returned in CCM search are held by 1 WRL only. (8769 records produced as result in CCM tool, compared with input file of 8817 records from GG report.) (-48 from GG total)	95% of records returned in CCM search are held by 1 WRL only

Headline Information.

From the results listed in Table 1 we can note:

- **That there is little difference between the results produced in the initial Art GG reports and the results produced through manual Excel matching. (Manual matching figure is 4 - 5% lower than than the GG total.)**
- **There is only 1% difference between the manual Excel matching results for records with or without ISBNs.**

- It was thought there may be a greater discrepancy between the results for records with /without ISBN, with the assumption that the testing would be much more accurate against the records containing the ISBNs; this is not reflected in the results.
- Matching on the same title/author/edition in GG appears to have failed on the occasions when there are discrepancies in the records of individual libraries in the formatting of the author and/or the publication details. It also appears to fail to match through irregular use of punctuation - for example the use of square brackets or non standard abbreviations.
- **When reviewing the results produced through the CCM Tool, the figure for non ISBN art records held only by 1 WRL in the WRL group were in line with the manual checking figures (95% of records returned in CCM search held by 1 WRL only.)**
- **Running the list compiled of WRL Art (700) records through the CCM Tool, the CCM result is 5% different from the local testing total.**
- **From the CCM results produced, 90% were identified as held by 1 WRL only within the WRL group.**
- **Broadly speaking these results (to me) reflect the results produced through the Physics testing. The greatest discrepancy is between Art and Physics non ISBN CCM results (7%.)**

Comparison with Physics Results

Table 2

Method of testing	Records with ISBN		Records without ISBN	
	Art (700)	Physics (530)	Art (700)	Physics (530)
Greenglass total	8817 records	5357	6952 records	5794
Local testing total	95% of GG figure for records with ISBNs 8413 records (-404 from GG total)	96% of GG figure for records with ISBNs 5137 records (-220 from GG total)	96 % of GG figure for records without ISBN 6656 records (-296 from GG total)	98% of GG figure for records without ISBN 5704 records (-90 from GG total)
CCM Tool testing	90 % of records returned in CCM search are held by 1 WRL only.	89% of GG total (4734)	95% of records returned in CCM search are held by 1 WRL only	88% of GG total (4677)

	(8769 records produced as result in CCM tool, compared with input file of 8817 records from GG report.)	(-623 from GG total)		(-1117 from GG total)
--	---	----------------------	--	-----------------------

Comment:

One thing noted from GG is that though it compares holdings for duplication across libraries - i.e. between York and Leeds and Sheffield, it does not compare within a “home” library for duplication. So it does not edit out multiple copies of the same item if they are on different catalogue records. This is as you would logically expect - in that there may be very good reason they are on different records and are to be treated individually - i.e. Rare Books, Provenance, historically a separate/branch library etc. In addition there can be multi volume titles with individual catalogue records for each volume part.

However historic cataloguing practices with all of their vagaries and errors do impact as well, and increase the GG total.

There were examples of this “internal duplication” in the Non ISBN stock in this subject area (particularly relating to York Minster Library, and also Leeds (Brotherton.)

Action: is it worth running one more subject area at this point? North American history (970)?? (noted as back up subject area.)

Review of overall results (Art)

GreenGlass results:

Art (DDC 700) Unique to one WRL within the WRL group.

- Total with ISBN = 8817
- Total without ISBN = 6952

Local Environment Checking:

- With ISBN - identified as unique to one WRL out of the WRL group:
8413/8817 = 95%
- With ISBN - identified as held by 2 or more WRL out of the WRL group:

$$404/8817 = 5\%$$

- Without ISBN - identified as unique to one WRL out of the WRL group:
 $6656/6952 = 96\%$
- Without ISBN - identified as held by 2 or more WRL out of the WRL group:
 $294/6952 = 4\%$

CCM tool results:

(Using GG report as original source of data for input file to tool)

- With ISBN - identified as unique to one WRL out of the WRL group:
 $7926/8769 = (90\%)$
- With ISBN - identified as held by 2 or more WRL out of the WRL group:
 $843/8769 = (10\%)$
- Without ISBN - identified as unique to one WRL out of the WRL group:
 $6480/6845 = 95\%$
Without ISBN - identified as held by 2 or more WRL out of the WRL group:
 $365/6845 = 5\%$

RE 17.05.17

Updated 18.05.17 (14:30)

Instructions for headline data checking

This same process should be run twice for each subject area:

- to identify stock held in (WR overlap =2)
- to identify (WR overlap = 3)

Subject area	Dewey Number
Maths	510
Education	370
Chemistry	540
Physics	530
French Literature	840
Psychology	150
Linguistics	410

Greenglass process

- Produce GG list in “home” library section of GG.
e.g. Specific Dewey ranges + Same edition + (WRL overlap =2) or (WR overlap =3)
- Name and save file.
- Export lists to Excel.
- Filter out records without ISBNs.
- Remove duplicate entries of ISBN Excel – select **Data tab**, then **Remove Duplicates- Expand Selection - Unselect All – ISBN.**)
- Add total number of records remaining in Excel into results table*
- Export data from Excel into Notepad++ and format appropriately:

[Formatting Bibliographic Record Number to enter as a “Batch” search in the CCM Tool, using Notepad ++

- *Copy list of ISBN from Excel into a new Notepad ++ file.*
- *Save file*

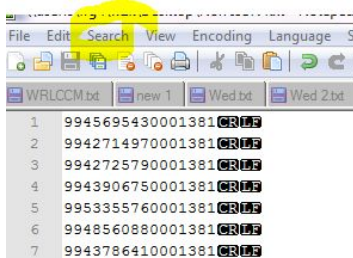
It may look something like this:

```

1 9945695430001381CRUS
2 9942714970001381CRUS
3 9942725790001381CRUS
4 9943906750001381CRUS
5 9953355760001381CRUS
6 9948560880001381CRUS
7 9943786410001381CRUS
8 9948004970001381CRUS
9 9954561480001381CRUS
10 9942676160001381CRUS
11 9947346140001381CRUS
12 9946196880001381CRUS
13 9948437300001381CRUS
14 9942789510001381CRUS
15 9944470390001381CRUS
16 9946444590001381CRUS
17 9948162830001381CRUS

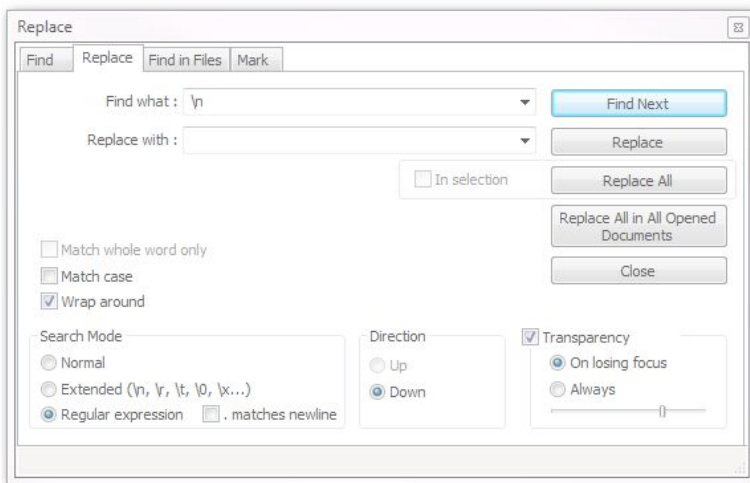
```

- Check that there are no column headers which need to be removed.
- Select **Search** on the tool bar



- And then **Replace** from the drop down menu.

You should see the box below.



- Ensure that the cursor is at the very beginning of the file.
- In **Find What** option enter \n

Replace with (nothing – leave blank.)

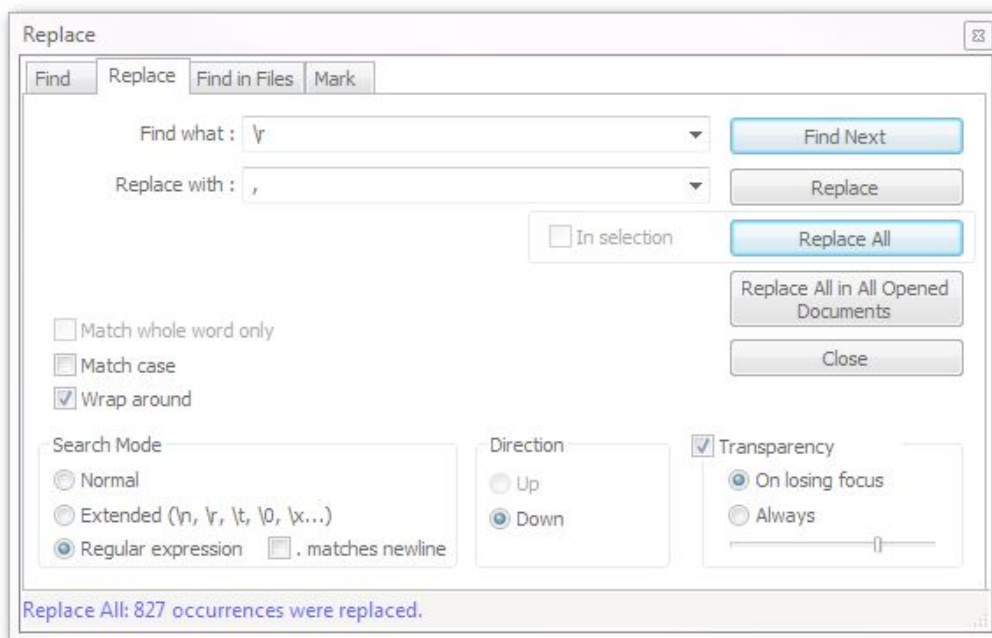
Click on **Replace All**.

- Then

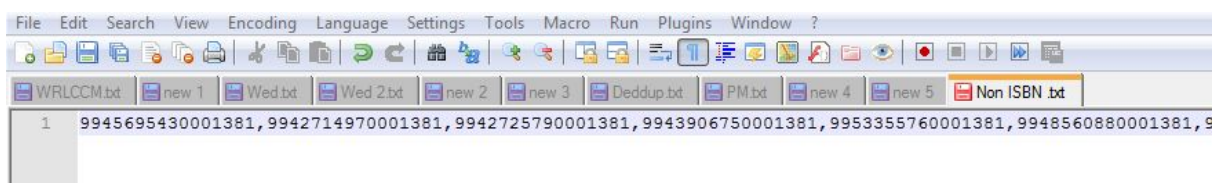
In **Find What** option enter \r

Replace with , (comma)

Click on **Replace All**.



- Click on **Close**.



List will appear horizontally – as shown above, with a comma between each number.

- Click on **Save** (under **File** dropdown menu.)]

Input file into CCM tool

Import file of local record numbers into the CCM tool:

- Go into **CCM Search**
- Select **Batch Search**
- From **Number Type** drop down menu select **ISBN**

Home » Search

Standard Number Search Batch Search Keyword Search

File Browse...

Number type ISBN

REFINE SEARCH

Library

- ☐ Warwick University
- ☐ Wellcome Library
- ☐ Wiener Library
- ☒ York Minster
- ☒ York University
- ☐ Zoological Society of London

Region

- ☒ All regions [Clear selections]
- ☐ North West England
- ☐ North East England
- ☐ Yorkshire & the Humber
- ☐ West Midlands
- ☐ East Midlands

Display records with All holding libraries

Deduplicate No deduplication

Search

Under **Browse** select the file you previously saved in Notepad ++
(It will look something like this.)



- Select and click on Open and the file will be brought into the CCM tool.
- Check **Number Type** is still ISBN
- In **Library option** select Home library (i.e. University of York or Leeds or Sheffield)
- (York should select YML (York Minster Library), UoY and NRM (National Railway Museum) libraries here.)
- Leave default as **No deduplication**.
- Select **Search**.

On completion of Batch Search in CCM tool:



- Scroll down to the bottom of the page for **Export Option: Items holding data**.

25. Astronauts and cosmonauts: biographical and statistical data: report, etc. [text]
 U.S.G.P.O. 1985
 Local record number: 9943113910001381

Held at:
 York University

Export: Full results set **Format:** Item holdings data .csv **Export** ?

- Click on **Export** and **Open** and then **Save** file.

Filtering in Excel

- Double check through the filtering process listed below that all the ISBN records listed have the home library listed as holding.
- Edit out any records which do not have the home library attached to them.

Filtering Processes in Excel

Note: in testing York has now realised that University of York, National Railway Museum and York Minster Libraries are separated out in Copac. Therefore when comparing holdings against York in Copac we need to produce figures for UoY, NRM and YML and then combine the figures together to give a York figure comparable to the GG figure.

	A	B	C	D
	standardNumber	briefRecord	numberOfHoldingLibraries	holdingLibraries
1	'9943906750001381'	3-J AND 6-J SYMBOLS. 1959		1 University of York Libraries
2	'9948004970001381'	A degree physics. Part 1. Arnold 1960 2nd ed.		1 University of York Libraries
3	'9947346140001381'	A PERSPECTIVE OF PHYSICS. Vol. 2. Selections from ... Comments on modern physics / introduced ... by Sir Rudolf Peierls. Gordon & Breach 1978		1 University of York Libraries
4	'9948437300001381'	A solution of the Navier- Stokes equations using a (u,v,p) formulation / [by] C. Greenough. 1992		1 University of York Libraries
5	'9944470390001381'	A text-book of physics: properties of matter; 8th ed. 1920		1 University of York Libraries
6	'9948162830001381'	A treatise on analytical statics, etc. U.P. 1896, 1892		1 University of York Libraries
7	'9948162750001381'	A treatise on statics, containing the fundamental principles of electrostatics and elasticity. U.P. 1880 2nd ed.		1 University of York Libraries
8	'9947356420001381'	A treatise on statics, with applications to physics. U.P. 1890 4th ed.		1 University of York Libraries
9	'9954226620001381'	The ABC of relativity / [by] Bertrand Russell, edited by Felix Pirani. G. Allen & Unwin [1958] Rev. ed.		National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Royal Society Library; The London Library; Trinity College Dublin Library; University of Birmingham Libraries; University of Cambridge Libraries; University of Glasgow Libraries; University of Liverpool Libraries; University of Oxford Libraries; University of Reading Library; University of Sheffield Libraries; University of Warwick Libraries; University of York Libraries

- Highlight **Holdings Library** column (column D in the example above.)
- In Excel tool bar select **Home/Editing/Sort & filter/Filter**.
- Click on arrow showing at the top of the selected column.
- In search box enter home library (as entered in Copac – i.e. - “University of York Libraries”).

(This should be all the titles in the list.)

- Press **OK**.
- A list of all titles listed as held by home library in Copac will show. Edit out any records which are not held by the Home library.
- In column E – enter a header denoting the home library, and then enter a Y (or S or L) in each cell in the E column which shows a title held in York. (As below.)

	A	B	C	D	E
1	standardNumber	briefRecord	numberOfHolds	holdingLibraries	York
2	'9943906750001381'	3-J AND 6-J SYMBOLS. 1959	1	University of York Libraries	Y
3	'9948004970001381'	A degree physics. Part 1. Arnold 1960 2nd ed.	1	University of York Libraries	Y
4	'9947346140001381'	A PERSPECTIVE OF PHYSICS. Vol. 2. Selections from ... Comments on modern physics / introduced ... by Sir Rudolf Peierls. Gordon & Breach 1978	1	University of York Libraries	Y
5	'9948437300001381'	A solution of the Navier- Stokes equations using a (u,v,p) formulation / [by] C. Greenough. 1992	1	University of York Libraries	Y
6	'9944470390001381'	A text-book of physics: properties of matter; 8th ed. 1920	1	University of York Libraries	Y
7	'9948162830001381'	A treatise on analytical statics, etc. U.P. 1896, 1892	1	University of York Libraries	Y


To be sure that the York figure is correct you will need to search on University of York Libraries, National Railway Museum and York Minster Library in a similar manner.

Then to show the total number of items held in York, total the three York libraries together as shown below.

- In the column H enter:
"Count" as heading
- In the next cell down enter:
=CountA(E2, F2, G2) and fill down.

This will show how many York libraries show with the record.

- As long as one or more York library has a record showing, replace the numerical figure with a Y, as shown below (otherwise further processes will not work.)
- Check for Leeds and Sheffield holdings in a similar manner using the filter.

K2  =COUNTA(H2:I2)

	A	B	C	D	E	F	G	H	I	J	K	L
1	standardNumber	briefRecord	numberOfHolds	holdingLibraries	Y	NRM	YML	Yk total	L	S	York total L+S	
2	'08176330:3.1416 and		2	University of Liverpool Libraries; University of York Libraries	Y			Y			1	
3				British Library; Imperial College London Library; King's College London Library; National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Newcastle University Libraries; Queen's University Belfast; Trinity College Dublin Library; UCL Institute of Education Library; University of London; University of Bristol Libraries; University of Cambridge Libraries; University of Cambridge Libraries (Special Collections); University of Edinburgh Libraries; University of Glasgow Libraries; University of Liverpool Libraries; University of Manchester Libraries; University of Nottingham Libraries; University of Oxford Libraries; University of Sheffield Libraries; University of Southampton Libraries; University of Warwick Libraries; University of York Libraries								
4	'97807619:100 statist		22	University of Sheffield Libraries; University of Southampton Libraries; University of Warwick Libraries; University of York Libraries	Y			Y		S		2
5				British Library; King's College London Library; National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Newcastle University Libraries; Queen Mary University of London Library; Science Museum Library; Trinity College Dublin Library; University of Cambridge Libraries; University of Exeter Libraries; University of Leeds Libraries; University of Liverpool Libraries; University of Oxford Libraries; University of Southampton Libraries; University of York	Y			Y				2

- Then In the column K enter:

“Count” as heading

- In the next cell down enter:

=CountA(H2, I2, J2) and fill down.

- Select column headed **Count** and filter to identify how many records (according to Copac) are held by the home library and one other of the WRL (WRL=2)....or (WRL=3). Enter into Results table ***.

A	B	C	D	E	F	G	H	I	J	K
standardh	briefReco	numberO	holdingLibraries	Y	NRM	YML	Yk total	L	S	York total L+S
			British Library; Imperial College London Library; King's College London Library; National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Newcastle University Libraries; Queen's University Belfast; Trinity College Dublin Library; UCL Institute of Education Library; University of London; University of Bristol Libraries; University of Cambridge Libraries; University of Cambridge Libraries (Special Collections); University of Edinburgh Libraries; University of Glasgow Libraries; University of Liverpool Libraries; University of Manchester Libraries; University of Nottingham Libraries; University of Oxford Libraries;							
'97807619	100 statist	22	University of Sheffield Libraries; University of Southampton Libraries; University of Warwick Libraries; University of York Libraries	Y			Y		S	2
			British Library; King's College London Library; National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Newcastle University Libraries; Queen Mary University of London Library; Science Museum Library; Trinity College Dublin Library; University of Cambridge Libraries; University of Exeter Libraries; University of Leeds Libraries; University of Liverpool Libraries; University of Oxford Libraries; University of Southampton Libraries; University of York							
'04715777	200% of n	15	Libraries	Y			Y	L		2
			British Library; Durham University Libraries; National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Newcastle University Libraries; Trinity College Dublin Library; University of Bristol Libraries; University of Cambridge Libraries; University of Leeds Libraries; University of Nottingham Libraries; University of Oxford Libraries; University of Reading Library; University of Southampton Libraries; University of Sussex Library; University of Warwick Libraries;							
'97801374	A first cou	16	University of York Libraries	Y			Y	L		2
			British Library; Cardiff University Libraries; Durham University Libraries; Imperial College London Library; National Library of Scotland; National Library of Wales / Llyfrgell Genedlaethol Cymru; Newcastle University Libraries; Queen Mary University of London Library; Trinity College Dublin Library; University College London Library; University of Aberdeen Libraries; University of Birmingham Libraries; University of Bristol Libraries; University of Cambridge Libraries; University of Edinburgh Libraries; University of Exeter Libraries; University of Glasgow Libraries; University of Leeds Libraries; University of Nottingham Libraries; University of Oxford Libraries; University of Southampton Libraries; University of St Andrews Libraries; University of Sussex Library; University of Warwick Libraries; University of							
'97804123	A handbo	25	York Libraries	Y			Y	L		2
			Cardiff University Libraries; King's College London Library; Newcastle University Libraries; University College London Library; University of Leeds Libraries; University of Liverpool Libraries; University of Manchester Libraries; University of Reading Library; University of Southampton Libraries; University of Sussex Library;							
'97804127	A handbo	12	University of Warwick Libraries; University of York Libraries	Y			Y	L		2

See below to record results.

Results: York

Home Library +1 (WR overlap = 2)

Dewey Number	Subject Area	*Number of Records in GG with ISBN (deduplicated) entered into CCM Tool	**Number of records returned by CCM tool	***Number of records identified in CCM tool as held in Home library +1
510	Maths	1864	1824	1372
370	Education	2487	2522	1910
540	Chemistry	860	734	559
530	Physics	970	836	625
840	French Literature	821	825	626
150	Psychology	1620	1626	1245
410	Linguistics	977	996	782

Home Library +2 (WR overlap = 3)

Dewey Number	Subject Area	*Number of Records in GG with ISBN (deduplicated)	**Number of records returned by CCM tool	***Number of records identified in CCM tool as held in Home library +2
510	Maths	1396	1377	1161
370	Education	2923	2952	2615
540	Chemistry	489	414	337
530	Physics	660	552	462
840	French Literature	512	518	440
150	Psychology	1087	1094	976
410	Linguistics	906	918	851

Results: Sheffield

Home Library +1 (WR overlap = 2)

Dewey Number	Subject Area	*Number of Records in GG with ISBN (deduplicated) entered into CCM Tool	**Number of records returned by CCM tool	***Number of records identified in CCM tool as held in Home library +1
510	Maths	3276	3291	2571
370	Education	6585	6634	5501
540	Chemistry	1239	1245	916
530	Physics	1718	1725	1316
840	French Literature	2262	2280	1835
150	Psychology	1360	1377	1052
410	Linguistics	669	672	536

Home Library +2 (WR overlap = 3)

Dewey Number	Subject Area	*Number of Records in GG with ISBN (deduplicated)	**Number of records returned by CCM tool	***Number of records identified in CCM tool as held in Home library +2
510	Maths	1433	1453	1136
370	Education	3698	3750	3200
540	Chemistry	534	537	334
530	Physics	823	824	507
840	French Literature	580	586	475
150	Psychology	938	956	832
410	Linguistics	672 ***	679	605
		*** Yes, there are more wrl3 than wrl2 for linguistics		

Results: Leeds

Leeds +1 (WR overlap = 2)

Dewey Number	Subject area	Number of records in GG with ISBN (De-duplicated)	Number of record returned by CCM tool	Number of records identified in CCM tool as held in Leeds +1
510	Maths	3500	3284	2694
370	Education	9764	9863	8152
540	Chemistry	1565	1494	1122
530	Physics	1771	1652	1307
840	French Literature	2282	2308	1872
150	Psychology	2108	2133	1688
410	Linguistics	1245	1259	1020

Leeds +2 (WR overlap = 3)

Dewey Number	Subject area	Number of records in GG with ISBN (De-duplicated)	Number of record returned by CCM tool	Number of records identified in CCM tool as held in Leeds +2
510	Maths	1371	1319	1099
370	Education	4232	4287	3690
540	Chemistry	511	484	332

530	Physics	724	660	466
840	French Literature	521	535	433
150	Psychology	1025	1045	895
410	Linguistics	685	694	614

Physics ISBN Testing

05.06.17

Questions 1:

the difference in totals between the number of records entered into the CCM tool (from an original GG sourced list), and the number of results which are produced as a result.

For example – 100 record numbers may have been imported to the tool, but results are produced for only 80. Which of the original records are not showing in the CCM results, and why is that. Also to look at the records which differed from GG in the manual spreadsheet.

Are the different methods (spreadsheet and CCM tool) presenting the same anomalies or different ones? That might possibly give us some insight into CCM matching.

York.

Physics (Dewey 530) WRL = 1

Total = 2659 records.

1530 with ISBN 1129 without ISBN

- Deduplicate lists

1139 with ISBN 883 without ISBN (deduplicated by Bib number)

Looking at ISBN list:

1139 records entered into CCM tool

982 records exported from CCM tool

Within the 982 records exported from CCM tool, 6 records are duplicated twice each (discrepancies in Metadata.)

When duplicates are edited out = 976 unique records.

Looking at a sample of 133 records produced by GG, 6 records are not listed in the CCM Tool report.

On checking – all discrepancies related to stock held in the External store and not list currently on Copac.

Looking at Non ISBN list:

GG deduplicated list (by Bib number) = 883 records without ISBN

883 entered into CCM Tool

581 records returned from CCM Tool

1 duplicated record found in list by title/publication date (on two separate bib records.)

Total number of unique records produced by CCM Tool = 580

Looking at a sample of 100 records produced by GG, 9 items were not list on the CCM tool report.

On checking – all discrepancies related to stock held in the External store and not list currently on Copac.

York Testing Update June 2017 - Art Stock

One of the tasks was to look at the differences in the numbers of records when the GreenGlass (GG) lists were imported into the CCM tool and the results returned.

[For physics see](#)

Example 1

Art (Dewey block 700-710). Items GG identified as unique (WR = 1)

Combined list of Leeds, Sheffield and York ISBNs listed as unique

8817 records imported into CCM tool with 8769 returned

Where are the missing records?

Looking at the combined list of 8817 records there are duplicate ISBNs in there. (There are records which GG has failed to match and have listed as unique. These same records appear in two or three of the library's own lists). When you de-duplicate the list on ISBN you get 8614 unique ISBNs.)

Re-importing the list of 8614 ISBNs into CCM returns the same number of results as before = 8769.

There are no records being 'lost' between GG and CCM, the discrepancy can be explained by duplication in the import file. In fact there are more results returned than imported.

CCM has in some cases returned multiple results for a single ISBN. This occurs when there are two or more separate COPAC master records with that ISBN.

Examples:

9780821206928

Sheffield is on different COPAC record to Leeds and York. 1 ISBN imported returns 2 results

<< <












Held At:  Google Preview
Exeter University 

Held At:  [Google Preview](#)
 Manchester University 
 Queen's University Belfast 
 Sheffield University 
 Warwick University 

Held At:  
Birmingham University 
British Library 
Edinburgh University 
Glasgow University 







Leeds and Sheffield copies on different records in COPAC.

Held At:  Google Preview
Sheffield University 
Tate Library (Tate Britain) 

Held At:           

9780300152661

York has two bib records with this item - one for print and one for electronic. These holdings have been attached to two separate records in COPAC. So one ISBN imported into CCM finds two matches in COPAC.

<p>Stepping-stones : discovering the cave artists of the Dordogne / Christine Desdemaines-Hugon ; foreword by Ian Tattersall. Desdemaines-Hugon, Christine 1946- New Haven : Yale University Press 2010 Printed</p>		<p>Held At:  British Museum</p>
<p>Stepping-stones : a journey through the Ice Age caves of the Dordogne / Christine Desdemaines-Hugon ; foreword by Ian Tattersall. Desdemaines-Hugon, Christine 1946- New Haven [Conn.] : Yale University Press c2010 Printed Online Printed (14) Online (3)</p>		<p>Held At:  British Library Cambridge University Leicester University Manchester University National Library of Scotland National Library of Wales Oxford University St Andrews University Trinity College Dublin University College London Wellcome Library York University 7 Fewer ...</p>
<p>Stepping-stones : a journey through the Ice Age caves of the Dordogne / Christine Desdemaines-Hugon ; foreword by Ian Tattersall. Desdemaines-Hugon, Christine 1946- New Haven [Conn.] : Yale University Press c2010 Online Online (6)</p>		<p>Held At:  Aberdeen University Liverpool University Queen's University Belfast Reading University Southampton University York University</p>

0140560041

This has produced two results in the CCM export. This ISBN produces multiple results in COPAC, including one for the 1953 published edition.

COPAC uses matches each institution's individual records to a master record, but retains the data from each institution's record. In this case one institution has erroneously included an ISBN on the 1953 edition, and COPAC has retained this information, creating another match.

[Art and architecture in France 1500-1700.](#)

Blunt, Anthony.

London : Penguin Books 1980

Printed

Held At: [British Library](#)

[Art and architecture in France 1500 to 1700 / \[by\] Anthony Blunt.](#)

Blunt, Anthony Frederick 1907-1983.

4th ed.

Harmondsworth : Penguin 1980

Printed

Held At: [Liverpool University](#)

[Art and architecture in France, 1500 to 1700: 4th ed. / \[By Blunt, Anthony.\]](#)

Blunt, Anthony.

Harmondsworth, Penguin Books, 1980

Printed

Held At: [York University](#)

[Art and architecture in France, 1500 to 1700 / by Anthony Blunt.](#)

Blunt, Anthony 1907-1983.

2nd ed.

Harmondsworth : Penguin 1970

Printed

Printed (24)

Held At: [Aberdeen University](#)
[Birmingham University](#)
[Bristol University](#)
[British Library](#)
[Cambridge University](#)
 19 More ...

[Art and architecture in France 1500 to 1700 / Anthony Blunt.](#)

Blunt, Anthony 1907-1983.

London : Penguin 1953

Printed

Printed (32)

Held At: [Bristol University](#)
[British Library](#)
[British School at Rome](#)
[Cambridge University](#)
[Cardiff University](#)
 26 More ...

[Art and architecture in France 1500 to 1700 / Anthony Blunt.](#)

Blunt, Anthony 1907-1983.

4th ed.

Harmondsworth : Penguin 1980, c1981

Printed

Printed (24)

Held At: [Birmingham University](#)
[British Library](#)
[Courtauld Institute of Art](#)
[Durham University](#)
[Edinburgh University](#)
 15 More ...

[Art and architecture in France, 1500 to 1700. / \[By Blunt, Anthony 1907-1983.\]](#)

Author Blunt, Anthony 1907-1983.

Series The Pelican history of art ; Z4
Pelican history of art ; Z4.

Published London ; Baltimore : Penguin Books [1953]

Physical description xvii, 312 p. : illus., plates, plans. ; 27 cm.

Notes Bibliography: p. 291-296.

Genre Bibliography

Biography

Illustrated

Format Printed

Held At: [Cambridge University](#)

[Art and architecture in France, 1500 to 1700 / by Anthony Blunt.](#)

Author Blunt, Anthony 1907-1983.

Series Pelican history of art ; Z4
Pelican history of art ; Z4

Published London : Penguin Books 1953

Physical description 312p.

ISBN **0140560041**

Genre Illustrated

Format Printed

Held At: [Cardiff University](#)

[Art and architecture in France, 1500 to 1700 / Anthony Blunt.](#)

Author Blunt, Anthony 1907-1983

Series The Pelican history of art ; 4.

Published London : Penguin 1953

Physical description xvii, 312 p., 192 p. of plates : ill., plans ; 27 cm.

Notes Includes bibliography and index.

Genre Illustrated

Format Printed

Held At: [Courtauld Institute of Art](#)

Example 3

York's list of unique non-ISBNS for Art.

2325 in GG list. Once de-duplicated on bib record number this gave 1954. Once imported into CCM 1930 records were returned.

Again, we believe this is a sensible result. The slight discrepancy is probably down to internal duplication - i.e. we have multiple bib records for the same title). If COPAC for example, has loaded a York Minster copy and a University copy onto the same master record, then two entries from the GG list will only return one result in CCM.

Some examples:

TRAGEDY AND THE PARADOX OF THE FORTUNATE FALL 1953

York has two bib records for this title. CCM would only return one result.

The symbolism of churches and church ornaments: a translation of the first book of the Rationale divinatorum officiorum 1843

York has three bib records for this title. CCM returns two results

Conclusion

From our point of view, testing the stock GG says is unique for art, the results we are getting out of CCM are not too problematic. There are discrepancies in the number of records imported into CCM and the results that come out. However we think there are duplicated records that we are putting into CCM, which can account for the 'loss of records'. On the other side, CCM sometimes returns multiple results for one ISBN imported - due to there being duplicated records in COPAC which have not matched.

Testing for Art WR = 3

Total number of records exported by GG before de-duplication

895 ISBNs returned from GG for which WR = 3

608 once de-duplicated (on ISBN)

608 ISBNs imported into CCM returns 618 results (with only York University, York Minster and NRM selected)

41 = 1 library

26 = 2 libraries

551 = 3 libraries

I also ran this again, but this time selected York, YM, NRM as well as Leeds and Sheffield in the CCM tool. This returned 687 results. The difference is that it picks up multiple records with the same ISBN, which have been loaded onto different COPAC master records.

For example: 9780500202418

Leeds and York on one record, Sheffield is on another.

Also tried importing the corresponding bib record numbers for this list of ISBNs. This returned 606 results from 608 imported.

Non-ISBN:

217 records

166 after de-duplication on bib number

CCM returns 159 results

There is a small amount of duplication on the input file which may account for the difference between 166 and 159

1 Library = 62

2 libraries = 16

3 libraries = 81

Conclusion

Again, we don't see any records going missing from the GG list imported into CCM. Differences can probably be accounted for by the reasons given above.

When we use the GG lists of WR = 3, it's clear that CCM doesn't detect that all of these are WR = 3. So the differences in matching work both ways. Some items GG fails to detect duplication, but CCM does, but the opposite is also true.

Greenglass lists WR = 3

Why are the WR = 3 lists for a particular subject not identical across the three libraries for a particular subject area? GG exports a list of items (each copy or volume on a separate row), so need to de-duplicate these to get a list of titles.

Possible reasons:

1. Internal duplication. For example two separate bib records for the same title. York does not share bib records with the York Minster Library, so there is some internal duplication. GG does not do any internal matching or deduplication. If York had 2 bib records for the same title, both of these could match to single records at Leeds and Sheffield and both would appear in York's WR = 3 list.
2. Cataloguing differences, particularly for multi-volume sets. Sheffield catalogues each volume on a separate record, whereas York (and Leeds?) largely does not. So for a 3 volume set Sheffield may have 3 occurrences on a list of WR = 3 but York may only have 1.
3. ISBNs. We split the lists up exported from GG by records which have an ISBN and those that don't. There may be records on the WR = 3 lists for which one library has an ISBN and the others don't. They may still match and appear on WR = 3, but will appear on our ISBN / non-ISBN lists accordingly.
4. Dewey numbers. GG takes the Dewey number from the bib record, unless no Dewey is present. If non present, GG will assign one. Sheffield uses Dewey numbers, but Leeds and York don't. However in some cases there is a Dewey number in the bib record that has been downloaded by the cataloguer. York doesn't delete these, so a significant number of our records may have them. There are cases where the Dewey number in the bib record is different to the one assigned by GG. As we've been looking at lists defined by Dewey ranges, this may mean that records for the same title may appear on different lists at different libraries.

Examples: From the WR = 3 lists of dewey range 700-710.

The relevance of the beautiful and other essays / Hans-Georg Gadamer ; translated by Nicholas Walker ; edited with an introduction by Robert Bernasconi. 1986

This appears on Leeds' list of WR = 3 (for art 700-710) but not York's. York's bib record has a dewey number 111.85 (so will appear on the list for that Dewey range)

The sculptor's workshop : tradition and theory from the Renaissance to the present / Rudolf Wittkower. 1974

On Leeds' list of WR = 3. York bib record has dewey number 731.4

Further details of our testing

In the same way that we looked at WRL=1 for each library we wanted to explore what results WRL=3 in GreenGlass would produce. Subsequent testing by all three WRLs showed that testing for an identical Dewey range with search criteria (WRL=3) did not produce an identical results for each of the WRL, as our initial assumption had been.

To understand this, analysis was completed on a GreenGlass Art (Dewey 700 - 710) report, which resulted in the totals for each library as given below. This shows that the totals held by each of the WRLs were similar but not identical.

Leeds	614
Sheffield	562
York	608
Total	1784

Summary of the analysis of the combined data from the table above

Key	Legend	Count	Titles	Workings
1	Leeds, Sheffield & York ISBN List matches	1284	428	Divided by 3 Libraries
2	Not on Leeds ISBN List	58	29	Divided by 2 Libraries
3	Not on Sheffield ISBN List	228	114	Divided by 2 Libraries
4	Not on York ISBN List	72	36	Divided by 2 Libraries
5	Only on Leeds ISBN List	39	39	
6	Only on Sheffield ISBN List	64	64	
7	Only on York ISBN List	25	25	
8	Queries	14	6	Refer to 6 titles
	Total	1784	741	

On further investigation it was found that the titles which did not appear on Leeds' ISBN list were indeed held by Leeds. The reasons for this are:

1. They were on the Leeds Non-ISBN 700 list
2. They contained a non-700 Dewey number in their catalogue record
3. They had been assigned a different Dewey number by GreenGlass

Consequently GreenGlass has correctly identified titles as being WR=3 which is useful for **stock checking**.

However, from Leeds' perspective, and from that of any other non-Dewey using library, the use of Dewey as a means of interrogating the data to **profile** stock has its limitations because it does not reflect the reality of the collection on the shelves at the home library.(JE)

Our initial assumption was that it should show the same total for each of the libraries – but with further exploration ...

"I had a look at the York list of art and physics for White Rose = 3 and saw a few things which may partially explain the discrepancies. Firstly, GreenGlass exports a list of items, so the numbers before de-duplication will not necessarily match up, as different libraries may have different numbers of copies.

We have 166 non ISBNs for art that are WR=3 once the list has been de-duplicated on Bib number (Leeds have 126 by comparison). It's clear that within our list we have duplicated titles that are catalogued on different bib records (we have titles that are held at both the main library and York Minster Library for example). GreenGlass doesn't do any internal deduplication or matching and because they are on different bib records, they are considered as separate titles. Even if Leeds and Sheffield only have 1 copy (and bib record) each for a title, both of our copies have matched across the WR and appear as on our list of WR=3. This knocks off 13 of our list and gives us 153 which is still higher than Leeds, but I think it shows we might not all get the same results for WR = 3.

Another thought that occurred that might be relevant particularly for ISBN results. For multi-volume sets, we generally have catalogued those on 1 bib record, whilst Sheffield has catalogued quite a few of these on separate bib records? A 3 vol set for example would return 3 results on Sheffield's list and only 1 on ours (provided the matching process thinks they are all duplicates) A couple of examples from our list of physics WR = 3 results

PROBLEMS IN UNDERGRADUATE PHYSICS 1965 (3 vol) is one 1 bib record at York, but on 3 at Sheffield

Twentieth century physics / edited by Laurie M. Brown, Abraham Pais, Sir Brian Pippard 1995 (3 vol) is one 1 bib record at York (with all the ISBNs, but on 3 at Sheffield (with separate ISBNs) I'd be interested to know if the Twentieth Century Physics example appears on Sheffield's list of WR = 3. GreenGlass is taking each one of our bib records and trying to find matches. If we have duplicate bib records within our own catalogue, or cataloguing differences (Such as multi vol sets on either 1 bib or separate bibles) it will return different numbers of results for each library, or (maybe even different results?) My interpretation would be that we don't necessarily have to try and get the WR=3 lists to match up exactly. If we do any comparison we would need to deduplicate our list (and in a consistent manner - probably on bib number rather than ISBN)" Email from MW York

Differences WRL have encountered in testing that might have affected matching in OCLC, GreenGlass or Copac



Typology of metadata issues

ISBNs

- ISBNs for different editions within same record

- Common practice to add e-book ISBNs to print records (& vice versa) could be problematic for matching

- Presence of qualifiers (pbk) / (hbk) following ISBN

- 13- / 10- digit ISBNs

Differences in name entries

- e.g. Oskamp, Stuart, 1930- (Sheffield) & Oskamp, Stuart (York)

Differences in titles

- Multi volume works catalogued by series title or individual vols

- Punctuation e.g. “Unemployment, 1920-1923” & “Unemployment 1920-1923” not matched

- Titles lacking statement of responsibility e.g. Marcellus Laroon / by Robert Raines (Leeds) & Marcellus Laroon (York)

- Additional names added to statement of responsibility e.g. translated by ...,

- Titles in capitals (York)

Presence of diacritics, symbols & abbreviations

- York used [] in titles

Differences in Publication places, publishers & dates

- Use of more than one place of publication

- Country designator included in one record but not another

- [s.n.] used in one record, when place recorded in other

- Different UK /US publishers for same title

- Publication date discrepancies

Differences in recorded size

- 21cm / 24cm - **why use size as a match criteria?**

- Pagination - **do differences in page numbers result in poor matching? Do we have examples?**

Series

- More than one series title recorded in a record

- Series titles recorded in 440 tag or 830 tag

Other issues

Print and ‘e’ recorded on same record