Helen K. R. Williams
Retrospective authority control

Article (Accepted version)
(Refereed)

Retrospective authority control

We began considering an authority control project at the LSE Library at the beginning of 2006. By this time the loading of all retro-con records was complete and an authority control group was convened in order to assess user needs. Authority control procedures had varied somewhat over the years, primarily because each library management system had offered different methods of verifying headings. As well as the legacy of system migrations, records had been imported from a variety of sources, and practices had become particularly unclear following the migration from Unicorn to Voyager in 2004. Only the Library of Congress Subject Headings file was purchased and so all new name authorities had to be authorised by manually importing the record from the Library of Congress. It became apparent that some staff were checking against existing entries to achieve consistency, others were importing Library of Congress records, while others were creating in-house authority records where none was available. This meant there were a number of variant headings in the catalogue.

The decision was taken to outsource a retrospective clean-up of all authority headings (subjects and series as well as names) from a company who could also provide a regular ongoing check of the catalogue. At the same time clear authority control guidelines were established for staff to ensure as few inaccurate headings in the catalogue as possible moving forwards.

Tenders were assessed and the project was awarded to Marcive. They would receive an electronic copy of our catalogue to verify all name, subject and series headings against Library of Congress authority files using automated processes. Bibliographic records would be amended to contain corrected headings. We would receive corrected records and new authority files for loading as well as reports of unmatched headings. This makes the process sound simple, but we discovered various complexities along the way.

Preparing to send the file to Marcive required some in-house planning. In particular we needed to be aware that any changes we made to our bibliographic records while the file was with Marcive would be overwritten when their corrected data was supplied. We therefore excluded order records from the data we sent so that we could continue to accession books during the project. These records were sent to Marcive for checking after the initial data clean, as part of our ongoing services. We kept a spreadsheet of existing bib records requiring changes during the course of the project so that they could be corrected afterwards. This built up into quite a considerable amount of work because the project took longer than we had anticipated.

In May 2007 we exported approximately one million bibliographic records to Marcive and two weeks later received a test file of 10,000 records for checking. We checked one in ten of these, which even so was a very time consuming process. We undertook such thorough checking aware that we had exported our entire catalogue to Marcive, and that the methods which had been used on this sample would be used on all our bibliographic records. Had we received records with errors or corrupted data back into our catalogue it would have a hugely detrimental impact on our users.

Our checking did reveal a number of queries which we submitted to Marcive. They provided a speedy and detailed response, though this indicated that we had higher expectations of the automated process than was actually achieveable. Fortunately the things we had hoped would be corrected through automated processing

appeared in accompanying error reports thus reassuring us that we would still be able to clean the catalogue to the degree we had originally intended, albeit that it would involve more staff time than we had hoped.

A few weeks later we received all our corrected data and accompanying authority records. Our IT department began loading three files into our test server. Unfortunately the test server suffered under the strain of so much data and we had to wait for Ex Libris to carry out a regeneration of the indexes before we could proceed. After this it was unsurprising to find that loading one million bibliographic records and 500,000 authority records into the live server was not without problems either. Having started the process IT estimated that loading the files and regenerating the keyword indexes after each one would take 30 days because it was such a slow process. One option was to take the live server offline, but downtime is inconvenient to staff and students alike so this was not a particularly practical option. The other option was to re-index in large batches at the end of the file loading. This meant that for about a week there would be inconsistencies in OPAC searches whereby the search facility used the old indexes but records contained new data. This seemed the most practical way forward, however, and 143 hours later all the data was loaded and the re-indexing completed just before Christmas 2007.

As 2008 began it was time to think about the ongoing processes Marcive would be providing for the Library. We send files of new records to Marcive on a monthly basis and they clean them and send them back along with any necessary authority records. Supplied with this is a report of anything unrecognised or with multiple matches and therefore requiring human intervention. This is worked through by a member of staff to tidy up the outstanding headings.

Running in parallel with this is a notification service whereby Marcive have a copy of all our authority headings and notify us if a relevant record is added to Library of Congress, or if there are changes to any of our existing authority records. These are dealt with through the Global Headings Change facility in Voyager. A member of staff can then approve, or disapprove, changes and this is applied to all related bibliographic records. This list is generated automatically by Marcive and we have found it requires manual intervention rather than automatically approving all changes, as some incorrect headings are suggested and we would not want these to be applied to all related bibliographic records. The Global Headings Change facility in Voyager was not as straightforward to use as we had hoped. We discovered eventually that this was due to a bug in the system meaning that we were unable to link new authority records to related bibliographic records. We had to wait for Ex Libris to resolve this and in the meantime had to stockpile the reports we were receiving from Marcive. Once this was resolved another bug meant we had 'orphaned headings' which would never clear from our list. After more work from Ex Libris this was sorted out as well and we were able to work on our backlog of reports.

In addition to these ongoing services from Marcive we still continue with our existing authority guidelines for in-house work. Authority work is far simpler with the item in hand as it prevents further work in terms of unmatched or possible duplicate headings which would often require retrieving the item from the shelf in order to correct the record at a later stage.

As well as embedding the ongoing services there was some tidying up work to be done on the headings Marcive had been unable to change and had notified us of in unrecognised headings reports. The personal names report alone had 250,000 lines and was so big it would not fit into one single excel spreadsheet. We employed a temporary member of staff and asked him to create a separate file of names

appearing on this report more than three times. This was on the assumption that we have some unusual material at LSE and that name authority records were less likely to be available from the Library of Congress for names occurring only once in our catalogue. As well as dealing with these multiple occurrences he was also able to work on the unrecognised subject headings report. We designated reports on corporate names, meeting names and series names as lower priority. Once the temp had completed high priority work we carried out a cost benefit analysis on the merits of completing the outstanding reports. Our sample testing suggested that authority records would not be available for over 90% of the remaining headings (those not already corrected by Marcive) so measurable benefits would be few in relation to the amount of work required in terms of time and cost.

The project could not have been completed without the hard work of those in the Bibliographic Services team who took part in testing data, our senior library assistant who oversaw the work of the temp and contributed in many other ways, and our IT team who persevered with the technological challenges.

We have been delighted with the result of all our hard work. The profile of authority control has been raised within Bibliographic services, which combined with our ongoing services from Marcive means we are in a strong position to keep the catalogue in good condition as we move forwards. As a result of the project our catalogue is now a great deal more consistent and has far less errors. In a library this size the catalogue is the primary way in which users identify the material we hold and so anything which makes that easier and improves accuracy is surely worth the effort involved.

A fuller write up of the project will be available in a forthcoming issue of CILIP's Update magazine.