Cleaning up the catalogue

The London School of Economics wanted to remove cataloguing inconsistencies but the scale of the task was huge, and outsourcing to a specialist bibliographic services company proved only a partial solution. **Helen Williams** explains why manual and automated processes were needed.

The library had good authority checking procedures in place for all newly added records, but we were aware that there had been a degree of inconsistency in the past.

IT IS ONLY possible to retrieve all the relevant records of a catalogue if there is a degree of standardisation in the way the catalogue is organised. To quote Michael Gorman, 'Cataloguing cannot exist without standardised access points, and authority control is the mechanism by which we achieve the necessary degree of standardisation'.

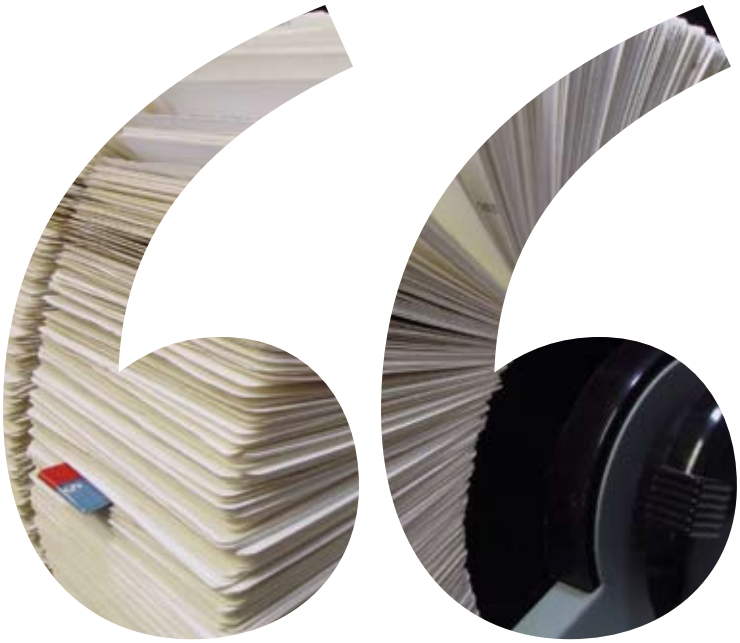
We began considering authority control at the London School of Economics Library at the beginning of 2006. The library had good authority checking procedures in place for all newly added records, whether scratch, vendor-supplied or downloaded, but we were aware that there had been a degree of inconsistency in the past.

A small working group estimated it would take 21,000 hours to complete the project in house, which wasn't feasible for us, so it recommended to senior management that the project was outsourced. The project was put out to tender in summer 2006. This required a company to undertake a full one-off automated check of all Name, Subject and Title headings in the catalogue (around 1m bibliographic records) against standard LC (Library of Congress) authority records and to correct any unauthorised headings. We asked for details of the proposed working method, strategy for checking headings, predicted success and error rates, treatment of headings not appearing in the authority file, likely duration of project and a quote for a regular check of the catalogue once the initial project had been completed, as well as a proposed working method for doing so.

Tenders were received from three companies and the project was finally awarded to Marcive. A sub-section of the initial working group then completed a detailed specification file sent by the company. While we waited for the necessary administration to take place at both ends, we considered how our in-house processes would be affected during the project. It was useful to be able to contact the UK reference sites given to us by Marcive to find out how other academic libraries using Voyager had worked with Marcive and the procedures they had put in place as a result of their projects.

In May 2007 we exported approximately 1m bibliographic records to Marcive, and waited for the results. We couldn't make any amendments to catalogue records on the system as these would be overwritten when the file was loaded back into our system. As our basic order records are on the catalogue, we decided not to send these so that we could continue to accession and process material throughout the project. We were then able to send these off to Marcive to be checked and cleaned in the first batch of records that went out as part of the authorities service once the initial clean-up had been completed. We also opted not to send any inter-library loan records (which do not appear on the front end of the catalogue), anything with an e-journals location (the updating of these records is already outsourced), or any miscellaneous test locations.

In approximately two weeks we received a test file of 10,000 records, which were checked by both an Assistant Librarian and Senior Library Assistant with cataloguing responsibilities, with help from a couple of other library assistants. We worked from printed lists of the test file catalogue records and checked them against records on our live catalogue. We made a decision to check one in 15 records; then another department in the library checked one in 10 of those. We checked that: indicators (numbers which affect the retrieval of bibliographic data from the catalogue) and punctuation had been correctly changed; the authority-controlled fields consisting of name, corporate name, subject and series entries corresponded to the correct authority records; old General Material Designations following title information had been correctly replaced with the newer phrase ('electronic resource'); and US/Gt Brit in subject headings had now been spelled out in full. We also made sure that the records looked correct and



We could not have done the project without the support of our IT department. We were fortunate that our IT representative has a cataloguing background and so had an invaluable understanding of both both technical and practical aspects.

that no corruption had taken place in the automated processes.

While there were no problems with the work Marcive had done for us, the checking did raise a number of other queries. Most of these were things which we expected would be changed as part of the checking/tidying up procedure. Marcive sent us a quick response, but we were slightly disappointed to find that most of its explanations were based on the fact this was an automated process. It highlighted that we had higher expectations of the process than could be met through automation. We established that the majority of these problems did at least appear on the accompanying error reports, reassuring us that we would still be able to clean the catalogue to the degree we had anticipated, albeit with more in-house input. We contacted Marcive and authorised the remainder of the work on our catalogue records.

Within a few weeks Marcive returned numerous files to us, comprising our entire catalogue. Aware that this data would overwrite our existing catalogue records, we arranged for our IT department to load some files on to our test server so that we could check files in a safe environment. As well as making sure we could see no problem with the records and the way in which the data loaded, we checked that authority fields were authorising on the system, thereby seeing that the records matched properly with the accompanying authority files Marcive had returned with the catalogue records.

Our in-depth checking was necessary, but reasonably time-consuming. It was not completely straightforward to load this much data back into the system. Our IT department found that it took approximately nine hours to load one file of catalogue records and that the system frequently crashed at the indexing stage. Clearly this was not a process that could be risked on the live server with a library full of users, so more thought was necessary.

By this time, there were a substantial number of catalogue records awaiting changes. I added these to a spreadsheet so that we could keep books going out to the shelves. The serials department is also heavily affected by this, as serials records need updating on a regular basis.

Once the data was finally loaded back into the system we had to work through the error reports supplied by Marcive which its automated process had not been able to change. They ran into hundreds of thousands of headings. For one thing, the automated process hadn't weeded out as many spelling mistakes as we had hoped. While Marcive has built up a file of common errors and mis-spellings, it obviously cannot cover everything.

We scoped out the work and took on a temporary member of staff for eight months to work on the reports we had designated as high priority. We were fortunate to have a committed temp who was interested in the work, and able to work quickly. Initially he cleared up around 7,000 unidentified subject headings, followed by some 4,000 unrecognised geographic main headings. The personal names report was much larger than this, and we wondered how to tackle it, given that there wasn't enough time for the temp to complete the entire report. Eventually we asked him to tackle only those names which appeared in the report more than three times. Our reasoning was that the one-off, or very low-use, headings, were less likely to have a LC authority record available; it made more sense to focus on records which we could correct. He completed around 6,000 names from this report. He also worked on approximately 16,000 records missing indicators and therefore not authorising correctly.

Finally he was able to begin work on some incorrect geographic subdivisions. This left us with some other reports from our medium- and low-priority lists. The medium-priority reports comprised the corporate names, multi-matches, remaining personal names and remaining geographic subdivisions reports, and the

> We had higher expectations of the process than could be met through automation.

Helen Williams is Assistant Librarian, Bibliographic Services, London School of Economic & Political Science Library (h.k.williams@lse.ac.uk).

low-priority reports covered the uniform titles, series and meeting names reports. These priorities were determined according to way we think our catalogue is mainly searched by users.

We opted to have two ongoing services offered by Marcive – the overnight authorities service and the notification service. This involves us exporting to Marcive, on a monthly basis, all the new records added to our catalogue. The data is cleaned by Marcive and returned to us, along with a multimatches report and an unrecognised main headings report which our Senior Library Assistant deals with. The notification service means that Marcive keeps an up-to-date list of all the subject, series, corporate and personal names used in our bibliographic records and can supply us with a new authority record if a LC authority has been newly created, or if changes are made to an existing authority record.

The first monthly file we exported to Marcive was much larger than our usual files as it contained all the records that had been added to the catalogue during the data clean itself and sorting out the problems of re-loading the data. When the monthly files are loaded, authority records which have been updated by Marcive are pushed into the Global Headings Change queue in Voyager. We check all these manually before authorising each change which then amends all the bib records linked to this authority. It is tempting to let all these changes go through without careful checking, but another institution had found some of these were not correct due to the automated processes – and we found the same. It took us a little while to understand how the GHC queue worked, and we had a few technical queries to work through, but after some initial bugs this now works smoothly and allows us to monitor Marcive's work.

We considered whether these ongoing services meant we no longer needed our stringent in-house authority validation checks at the cataloguing stage, but decided that we should continue with them. Authority work is far simpler with the item in hand; if there are unmatched or duplicate headings it could mean retrieving the item from the shelf to correct at a later stage.

We could not have done the project without the support of our IT department. We were fortunate that our IT representative has a cataloguing background and so had an invaluable understanding of both technical and practical aspects.

Although the project has been time-consuming, it has been worthwhile. Our catalogue is now is more consistent and has fewer errors, making retrieval more straightforward for users. In a library this size the catalogue is the primary way in which users identify our holdings. Our library catalogue continues to get a high score on the student satisfaction survey, which shows that the hours put into this project have been fruitful.