The institutional repository and the library catalogue: adventures in data conversion

Sally Rimmer, Cataloguer and Metadata Librarian, University of Derby

Having your institutional repository records integrated into your library catalogue is acknowledged to be a good thing. Who wouldn't want all their institution's research outputs made visible and searchable alongside its other collections? However, costs to outsource this work can be prohibitive and it is often put aside as a "nice to have" facility.

So it was at the University of Derby until circumstances changed. The appointment of a Repository Librarian, a newly created post, and the impact of REF 2014 led to a greater awareness of our repository UDORA (University of Derby Online Research Archive). There was a subsequent increase in submissions further enhanced through promotional events showcasing the service but the impetus to really consider integrating UDORA within our library catalogue (Capita PRISM) and hence our discovery service (EBSCO EDS known at Derby as Library Plus) finally came after an enquiry from an academic who was nonplussed as to why the UDORA records weren't in the catalogue. The time was right to start a project to see if it could be done in-house.

From the start we wanted it to be feasible and sustainable for one part time cataloguer to pull into their existing workload and knew that, for maximum effect, UDORA records would preferably be searchable as a discrete collection within the library catalogue.

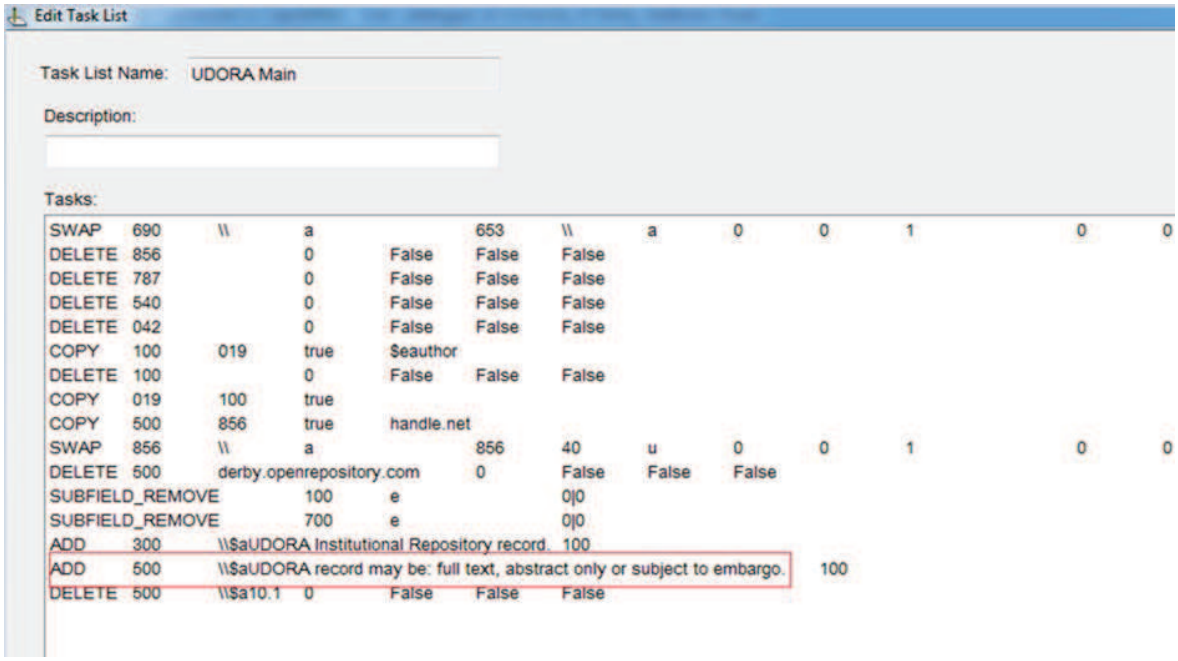
Preliminary work started by examining the metadata produced by our repository. Our repository is from Atmire[1] and the metadata scheme used is Dublin Core so the primary task was how to convert the Dublin Core to MARC data. There is a published crosswalk[2] and the first attempt consisted of downloading one record from UDORA with the aim to transpose the elements to MARC fields with a view to somehow automating this later. This was simple to do but of course with over 2,000 records in UDORA, and the ideas for somehow automating it later in short supply, manual retrospective transposition was not an option plus we had new submissions coming in all the time.

Further research led to OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) which is when the possible use of MarcEdit[3] came to mind. MarcEdit, freely available, allows you to manipulate your metadata easily. Previously MarcEdit had only been used at Derby for editing e-book records prior to upload into the catalogue but it also has an OAI-PMH harvesting facility in its toolkit. Serious upskilling in MarcEdit was done by means of Youtube videos[4] and a very comprehensive set of online guides produced by the University of Illinois[5] which provided the knowledge to set about harvesting, converting and editing our repository data into usable Marc.

---

1. http://www.openrepository.com/
2. https://www.loc.gov/marc/dccross.html
3. https://marcedit.reeset.net/
4. https://www.youtube.com/user/tpreese/videos
5. http://guides.library.illinois.edu/MarcEdit

We were then on to the concept proving stage. Marcedit allows you to download the whole repository, collections or a particular subset using a date range. The latter was chosen to produce a small test set. Marc came out but in a raw state which needed quite a bit of editing but this is quite simple as MarcEdit allows you to easily manipulate data by performing bulk editing using sequenced specific tasks. You can also save these sequenced tasks as task lists for repeated use. After trial and error our eventual task list was set at:



Some of the functions seen used here are swap, delete, copy, subfield remove, add. For instance we added a note to say that the record may be full text, abstract only or subject to embargo to clarify what the record is pointing to. We also deleted the relator subfields purely because in the first instance, for speed, we made them AACR2 records. Going forward we will use RDA and the RDA helper within MarcEdit to do so.

Our Library Management System (Capita Alto transitioning to Soprano) allows you to bulk import records to a particular profile which we used to import the test set. We set the profile up to enable us to run a management query which would provide the SQL needed to create a discrete UDORA collection in PRISM. We also needed a separate profile for our theses as we also wanted them to be picked up by the existing Theses and Dissertation collection in PRISM.

The concept was proved but how long would it take to do all the retrospective work and would it be feasible to fit new submissions into a regular upload schedule bearing in mind the time constraints of a part time cataloguer? Using MarcEdit to download the entire repository the sequenced tasks were applied with a few fields (001, 003, 008) added separately in bulk at the end. Using the MarcEdit split function we divided the approximately 2,300 records into batches of 100 to make it easier to manage. Each of these files were quickly checked and any errors amended. Each file took approximately 45 minutes to fully prepare and load including the manual tidying (which included coding subfields correctly and some series entry editing). The theses records were hived off to work on separately later as slightly different tasks needed to be applied. Over seven weeks the retrospective editing and importing of the 2,300 records took place as it fortunately coincided with a quiet period in the cataloguing year.

Work also began on formulating a process for monthly updates. MarcEdit does allow you to specify a date range to harvest from the repository but unfortunately in our instance it returned anything new or edited during the specified period.

However, it is relatively easy to tell by their repository handle number as to which are the most recent and you can sort records by field within MarcEdit.

Once we had finished loading all our existing repository records into the catalogue, and had a method for updating the new submissions, issues surrounding this upload became apparent. Specifically, the UDORA records were inconsistently skewing results in our discovery service and library catalogue. In the discovery service UDORA records would sometimes appear at the top of the results list ahead of full text subscriptions. This was problematic as the UDORA records may or may not lead to full text (depending on embargoes and other access restrictions). If the first link doesn't resolve to full text then there is always the risk of a student thinking that we don't have access when we might. Also we couldn't easily identify and only load repository records where we didn't already have a full text subscription and even if we could as subscriptions change quite frequently we couldn't keep amending the catalogue records to keep pace with them.

In our library catalogue (PRISM) a similar problem occurred. We have a results clustering facility enabled which groups together all formats and editions of the same title under the umbrella record of the latest version. Where the UDORA record was the latest version it could look like it was the only version available and if a student found that it didn't resolve to full text they might conclude that we didn't have full access to the work. Finally our attempts to create a discrete repository collection in our library catalogue were also thwarted by technical issues.

To solve the book clustering issue it was decided that we would purchase copies of books where our research community had had input and suppress the relevant UDORA records. This was a good news story to tell as well as solving this problem. Work on creating a discrete collection for UDORA items is still ongoing in collaboration with our IT services and library management system suppliers.

However, the discovery issue, once it became apparent that we couldn't fix it in the parameters of the discovery service, was more serious. Requests to see if the UDORA records could always be the last in a series of results was not possible as a subset of the catalogue could not be separated out from the rest of the catalogue records. So we took the decision to suppress all the repository records that were journal articles which left only the grey literature to be added to the catalogue and hence into our discovery service. We haven't given up on the idea of integrating the journal articles into our catalogue and will be working with our discovery service and repository providers as technology advances to progress this.

Positive outcomes for us include adding grey literature to the catalogue on a monthly basis and the future project of adding our e-theses. We are also investigating using MarcEdit to convert repository metadata into a format acceptable for use in an RSS feed on our university website.

The project raised some interesting, wider, discussion points in our library community. Library catalogues don't exist in isolation anymore. They are often connected to reading list services and discovery services which can be impacted, positively or adversely, by changes you make to your library catalogue. Then there is the nature of the catalogue itself, it traditionally points to full text and our repository records did not always resolve to full text. Whatever, the rich data we produce needs to work hard for us and tools such as MarcEdit can help us to this.

Biography

Sally Rimmer's role includes a remit to advise on the metadata aspects of the University of Derby's institutional repository (UDORA) and because of this she has become interested in converting and using data in multiple ways.
s.e.rimmer@derby.ac.uk