Harvesting repository (EPrints) records for use in an LMS (Aleph)

As with many things, making a small decision can be the thin end of the wedge (in a good way!) in terms of both creating work for yourself and making improvements. That was the way with this project but I think the overall benefits have been worth it.

In early 2011 the decision was made to include details of all the higher level research theses (PhD, MPhil and Masters by Research) in the Kingston University Research Repository (http://eprints.kingston.ac.uk). It was just the metadata as we've not yet gone down the electronic theses route. The repository records were to include elements such as supervisor's name and funder, elements not listed in the catalogue records. This meant each thesis (about 600) was checked by hand and the details entered into a spreadsheet. This spreadsheet was then uploaded to the repository by EPrints as part of a paid upgrade.

The process of going through the theses led us to thinking about the whole process of adding theses to stock. It became apparent that there was a certain degree of double-handling going on which meant theses were taking some time to be available. From participating in the British Library's EThOS scheme, we knew that getting the theses available quickly was important as they were often requested very soon after being added to EThOS. We already knew that the catalogue records were not as rich as those in the Repository as these were often catalogued quickly by a non-cataloguing member of staff. We began to wonder if, instead of detailed records being added to the repository then being added separately to the catalogue, we could take the information from the repository and import it directly into the library management system.

By this point we had moved to a new library management system (Aleph) in summer 2011. Looking in the manual it showed where text-based records could be imported but very little information was available on what format the text had to be in in order to display properly. I did some hunting around on the internet, at times it

felt like I was looking for a needle in a haystack! Finally I found a document from the University of Liege in Belgium (http://orbi.ulg.ac.be/handle/2268/494) which explained how to enter the bibliographic information in a spreadsheet in order to output it in a text file that was suitable for importing.

Meanwhile, our technical analyst was looking for a programming project to work on in conjunction with EPrints so we asked if there was a way to automatically fill the spreadsheet from the Repository. This meant I had to explain which parts of the record I wanted and in what format I wanted the text to appear in the spreadsheet. You can choose which elements to export from the Repository, for example, we don't take the 'Research Area' as it's not a Library of Congress subject heading. In the records on EThOS, however, 'Research Area' is incorporated as 'Keywords'.

A colleague of mine explained how to interpret the XML and how to write what I wanted it to look like so that a computer could read it in terms of spacing etc… In the end it worked out better than we'd hoped as a Perl script was written to output from the Repository straight to a Word file ready to convert to text for importing[1]. So now, once a search has been done in the repository, the results can be exported into Word and imported into Aleph via a text file.We've chosen not to go back and 'upgrade' the old theses, although we may consider scanning in abstracts of them all as and when time allows.

Of course, there are often different ways of doing the same thing. Since setting up the export I have found out I could have used MARCEdit to carry out this work. At the time I had considered this may be a possibility but I couldn't find the information and we had the support for creating the export through EPrints.

As with many projects, at times it felt like I'd opened a bit of a can of worms with this! When we chose to add a link to the EThOS record in the Repository we

[1] Information on writing an API can be found at (http://wiki.eprints.org/w/How_to_write_plugins)

found that some of our theses were not listed on EThOS. Many of our early theses were awarded by the Council for National Academic Awards (we're a former polytechnic) and some of these were incorrectly assigned to other institutions. Sara Gould and Heather Rosie at EThOS have been brilliant at working with us to correct these errors. Furthermore the process of adding new theses to EThOS has been greatly streamlined. Before this, title pages and abstracts were photocopied and sent by post but now the records can be exported directly to EThOS from the repository (they use the date stamp to 'pick up' records added/updated).

Overall though, despite the work involved in getting the project to this stage, there have been a lot of plus points. I've got better understanding of the workings of the repository. Also, I thought I was pretty good at being able to explain cataloguing but this really did test it, not only in explaining it to other people but how to write things in a way that a computer will understand and output what you want and I've learnt a lot. The main bonus, however, has been one that will benefit many people in that the catalogue (and EThOS) now more comprehensively covers what we have, with greater findability.

The slides of my talk, along with further information on creating a spreadsheet and text file can be found at http://eprints.kingston.ac.uk/23511/ (on our institutional repository). I'm happy for people to contact me to find out more out what we did if they are interested.

Katrina Clifford
Senior Information Advisor – Kingston University