Subject indexing in an institutional repository

Clare Playforth, Cataloguing Library Assistant at University of Sussex, student member of Society of Indexers

I've been a cataloguer for some years but have only just started training to become an indexer with the Society of Indexers. I can now see that there are many parallels between cataloguing and indexing and I am often expanding my knowledge of one activity through the other. The clearest example of a task in which the two areas are intertwined is when I classify theses in our institutional repository. Our current repository platform is EPrints using the Dublin Core Metadata Element Set. This allows us to assign subjects to research outputs so that they are indexed and available to users through access points in our discovery layer (Primo). I'm going to avoid discussion about the systems involved here and their interaction with each other and am going to focus on the details of this task and try to understand some of the benefits and flaws of the current workflow.

When a masters or a doctoral degree is awarded the cataloguing team are notified and we create an entry in the repository and upload the thesis manuscript applying any embargoes where necessary. Once the record has been created we classify the thesis by selecting terms from a subject tree built on Library of Congress authority records and structured using the Library of Congress Classification scheme (LCC). This is also the scheme we use for organising certain sections of our print collections on the shelves.

The top level headings in the subject tree are the main classes which you can see here in the Library of Congress Classification Outline. Our repository software displays a list of these and each heading or 'subject node' can either be selected by clicking 'add' or expanded by clicking '+' to show any further subject headings or 'children'. In this way the tree has a hierarchical structure that nests multiple levels of subheadings under the top level entries potentially allowing a very specific classification to be constructed in much the same way we would when creating a classmark for a print book. If, for example, the research is about Don DeLillo then we could start at P Language and Literature and progressively click through, expanding the headings until arriving at the required level of specificity.

P Language and Literature

PS American Literature,

PS0700 Individual authors,

PS3550 1961-2000,

PS3554.E4425 DeLillo, Don

This process can be repeated and multiple subjects can be selected in order to classify a thesis which spans subject areas. Our discovery layer will display the final term in the string (e.g. DeLillo, Don) minus the class mark, under the 'Subject' facet as an access point in the same way it would display the contents of a 6** field from a MARC record.

If we are unable to find a suitable classification for the thesis then we can add a new subject node as required. For example if we needed to classify our first PhD thesis about Madeline DeFrees we would need to create an entry for her in the tree as a child of the 1961-2000 node. Her subject would display above DeLillo as the LC has assigned her the call number of PS3554.E4 meaning she files before him (but after Charles Bukowski who is at PS3552.U4).

Areas of LCC such as this where author numbers are used lead to small sections within this index where the entries are filed chronologically then alphabetically like this…

PS3500 1900-1960

PS3550 1961-2000

PS3551.U77 Auster, Paul

PS3552.U4 Bukowski, Charles

PS3554.E4 DeFrees, Madeline

PS3554.E4425 DeLillo, Don

PS3563.C337 McCarthy, Cormac

PS3566.Y55 Pynchon, Thomas

PS3568.O855 Roth, Philip

…but in the main it is hierarchically organised and hence displays the bias of the LCC original creators. The structural problems with the hierarchy are embedded but we do have the choice to reject some of the more outdated and questionable LCC terms and codes here. If an entry is offensive or inappropriate or we decide it is not really in current usage then we have the option not to add it to our classification tree. Much like we adapt the LCC scheme to fit our own print book collections and shelving requirements we can adapt it for the repository if we deem it necessary. This is a plus point for an index that grows organically with a collection (as opposed to one built on imported data), we can pick and choose as we go.

Assigning subjects is not an easy task and requires careful analysis of the document. We always read through at least the abstract and try to work out what the thesis is about, in other words we try to understand what the metatopic is. This is a particular challenge for PhD theses because by their nature they contain new concepts that may never have been explored before. We use the LCC schedules and classes from existing items in our own collections to help us.

Although, as I said, we do have a choice about what we add to this index it is still necessarily created from a controlled vocabulary and with that comes some problems. The person doing the indexing has to know how the classification scheme works in order to assign their terms. It takes training and experience to become efficient at this task. Say we had a thesis about pubs, you'd have to be at least a little familiar with LCC to know that you might need to look under the Technology heading in order to find TX0950 Taverns, barrooms, saloons. This has connotations both for the indexer and the user because although the LC authority file might have linked an equivalence relationship between two concepts, this is not translated into our index and synonymous terms would need to be searched for individually by the user.

So why bother with this classification tree, why not have free text entry for subject terms? The repository does have an option for keywords to be entered into the records with the rest of the metadata about the item and you are able to specify them in the advanced search function of the repository itself. However when I tested a few of these it seems they are not indexed in our discovery layer and seeing as the majority of users are accessing our resources from our main Library Search pages we could say that these keywords do not aid discovery as well as terms from the controlled vocabulary of the subject tree. Using the LCC scheme and authority files also allows for disambiguation because qualifiers such as dates are available for otherwise identical entries (authors with the same name for example).

Another reason for assigning subjects from the classification tree is that we are able to classify theses by their theme and area of study. If we were to rely on keyword searches that simply return matches of words from the abstract or title it is not adequate because the metatopic of the thesis may not be mentioned by name. If it is merely implied or buried in specialist terminology then it may need teasing out and highlighting and it is the job of the cataloguer to do this.

Indexing the subjects of theses is a tiny piece in the puzzle of the changing scholarly communications scene and as this scene develops I expect we are going to find systems are better integrated and processes for metadata creation will become more automated. However if there's one thing I have learnt from my indexing training it's that the initial selection of subject terms cannot be satisfactorily automated. Even when we are able to create links and share metadata totally seamlessly, it is unlikely that there will be a way to replace the human intellectual input in the analysis of texts at the academic level. It is through the process of creating new subject nodes described above that we have gradually been able to expand the classification tree over the years, producing ever richer subject options for future cataloguers/indexers to use and add to, hence continually improving the discovery potential for the university's research outputs.

References

Batley, S., 2005. *Classification in theory and practice*. Oxford: Chandos Publishing.

Library of Congress. 2015. *Classification and Shelflisting*. [ONLINE] Available at: https://www.loc.gov/aba/cataloging/classification/. [Accessed 31 July 2018].

Library of Congress. 2016. *Library of Congress Authorities*. [ONLINE] Available at: https://authorities.loc.gov. [Accessed 31 July 2018].

Library of Congress. 2015. *Subject and Genre/Form Headings*. [ONLINE] Available at: https://www.loc.gov/aba/cataloging/subject/. [Accessed 31 July 2018].

Society of Indexers 2010, 2014 and 2016. *Training in indexing*, 4th edition. Sheffield: Society of Indexers.