Mirror mirror on the wall does your repository reflect it all?

Peter West and Timothy Miles-Board
EPrints Services
University of Southampton
Southampton, UK
pjw@ecs.soton.ac.uk tmb@ecs.soton.ac.uk

Introduction

The conference aims state that "One of the most important roles of repositories is to enable greater use and reuse of their contents". It stands to reason that the quality of the data held in a repository has an important influence on the ability to and desirability of using and reusing that data.

As repository service and support providers we work with many different institutions around the world. Our presentation will describe the journey that we are taking with our clients towards being able to answer questions such as: *Does my repository accurately reflect the published output of my institution?* In order words, how can we help repository administrators *validate* the completeness and accuracy of their repository holdings?

In attempting to answer these questions we have embarked on a process of enquiry, development and support that we hope will lead us to a set of tools and procedures for the repository community.

Community Engagement

In an effort to clarify the problems faced by repository owners we have engaged in a number of discussions with repository managers and administrators. Initial discussions have indicated that data validation is a real and pressing issue and that it would be desirable for the repository to provide further help and support for solving or at least alleviating the problem. Owners have indicated concerns regarding the relationship between the contents of a repository and the research outputs of an organisation. These include:

1. Does our repository content accurately reflect the published output of our institution?
2. Is our bibliographic metadata accurate and complete?
3. Are our publications correctly and unambiguously associated with the right authors, editors, contributors?

Our presentation will use case studies to focus on each of these questions in turn.

Case Study 1

Does our repository content accurately reflect the published output of our institution?
An interesting example of the need to answer this question came out of a repository that we built for the research and development division of a large multi-national company. We set up a repository to drive an internal approval workflow for all research that is to be published and/or presented externally. Before external publication/dissemination, researchers must submit their work to this repository where it is reviewed by supervisors, department heads and patent lawyers. Once the work passes approval the researcher is free to submit it to publishers and/or present the work outside of the institution.

Much more recently it has become apparent that the repository administrators are highly motivated to demonstrate that all known work published by the division can be linked back to a successful

approval in their repository. Instances where research is published without first being approved need to be identified quickly so that remedial action (in the form of user training) can be taken. We think that this is a good driver for the development of tools for comparing a set of repository holdings with a set of known published work that will have a wider applicability to the repository community.

We have designed a system whereby the repository administrator can collate a list of known published work (for example by searching publisher databases such as Scopus, Web of Knowledge and PubMed) and regularly upload that list to their repository. The repository runs a "publication match" process which attempts to match each externally published item to an approved item in the repository. The results of this process are presented to the repository administrator as follows:

- If an externally published item is matched to an approved item in the repository, the administrator can mark the repository item as "published" and merge the publisher's bibliographic metadata into the repository item.
- If an item is approved in the repository but has not been matched (after, say, 12 months of being approved) the administrator can mark the repository item as "not published".

Externally published items which are not matched to an approved item in the repository are highlighted - the administrator is therefore able to directly report on the total number of items which were approved for external publication, the number of items which actually went on to be published, and instances where a researcher may have published information that had not first been approved.

Case Study 2

Is our bibliographic metadata accurate and complete?
For one university the issue of accurate journal and publisher information had become an issue that was affecting the efficiency of both the repository's editorial team and its submitters. One of the key problems was the fact that submission to refereed journals is an important factor in the annual review of research outputs for that university i.e. this directly impacts funding allocation. Accurately identifying a journal and a publisher and having access to canonical data concerning the refereed status of that journal was increasingly important.

Initial attempts to alleviate this issue had relied on auto completion of journal names for submitters plus the collection of journal and publisher data by the editorial team. It was realised that the data collected by the editorial team could be utilised more effectively if it could be more tightly integrated into the submission process.

A database of journal information (JDB) was developed that is tightly integrated with the repository. Now a user retrieves journal and publisher data for an item they are submitting using an interactive dialog that leads them through a search of the JDB. Mechanisms are in place to allow users to search other external databases if no data is found in the JDB and in the worst case the user can manually enter the data. (Data that is manually entered or found in external databases during this process is added to the JDB automatically). Using this dialog submitters are providing better quality metadata to the editorial team and they are faced with one interactive dialog rather than a number of blank fields on a submission form.

Data integrity is an important issue for the JDB and different mechanisms have been put in place to check for duplicate entries and broken links. The duplicate detection checks for duplicated journal or publisher entries using source IDs, identical names, identical abbreviated names or ISSNs. There are also checks for similar journal or publisher names.

An experiment is in progress to build on the experience of developing the JDB. In this experiment we are attempting to harvest journal and publisher data from open access repositories. In essence this is an attempt to build on the combined efforts of all submitters who have provided journal and publisher data for the purpose of building a database of canonical journal and publisher data. The main issues identified thus far are the volume of data, duplicate management and achieving automated updates so that the data remains current.

Case Study 3

Are our publications correctly and unambiguously associated with the right authors, editors, contributors?
This is a problem faced by many repository owners. The solution adopted by one university was to leverage their single sign-on data. By utilising the techniques developed in Case Study 2 we were able to provide the ability to integrate external data sources into the submission process. Specifically, free-text input fields on a submission form were replaced with an interactive dialog that guides the submitter through the process of selecting the appropriate contributor and their contribution. In an interesting parallel to the development of the JDB it was found that it was possible to utilise the contributor data in additional ways, for example it is now possible to automatically allocate the affiliation of the item being submitted based upon the data retrieved about the contributor.

Looking forwards it is likely that organisations will wish to extend internal contributor disambiguation to external contributors. To meet this need we are in the process of extending the techniques developed in the case studies presented here to include external sources of contributor data such as the ORCID.

In general, where data accuracy is important, widgets that allow users to interactively search an authority list should replace free-text entry fields.

About EPrints Services

EPrints Services is a UK-based repository software developer and service provider. We help many repository administrators around the world to get the best from their EPrints repository. In 2013 we are focusing on three core challenges -- Validation (covered in this presentation), Reporting (*how can your repository help you report to your institution, funding bodies, government agencies?*) and Dataset management (*how can your repository present and disseminate your research data?*) -- and are working closely with our user community to meet these challenges.