sciecominfo
Nordic-Baltic Forum for Scientific Communication

DATA QUALITY ASSURANCE OF RESEARCH PUBLICATIONS - THE CASE OF MALMÖ UNIVERSITY INSTITUTIONAL REPOSITORY

Pablo Tapia Lagunas

Introduction

This article presents the results of a small scale study aimed at finding the most common changes and additions to items in Malmö University's institutional repository and to use this information to design an approach and strategy for the repository staff in relation to the researchers registering their publications. The article is a continuation of an earlier project at Malmö University with the aim to implement a new stage of quality assurance in the workflow of the repository.

Malmö University Electronic Publishing (MUEP) is Malmö University's open access repository for scholarly output. It is also the open archive for publication series published by Malmö University. MUEP is based on DSpace open source software.

MUEP has been the institutional repository at Malmö University since 2003. From publication year 2007 onwards, MUEP also forms the basis for local research assessment and aims at having full coverage of what is published. Approximately 20% of the research publications are freely accessible. In November 2010 Malmö University decided on a new open access policy[1].

In order to increase the quality of publication data registered by the academic staff, a project was initiated in March 2010 with the purpose of creating a third stage of quality assurance in the workflow of research publications in MUEP.

The first stage in the registration workflow is the registration process completed by the researcher: the registration of all necessary data and metadata regarding the publication.

The second stage is an accuracy check effectuated by the research officer at each faculty who primarily checks the affiliation of the author. Difficulties can appear when the publication is registered retroactively. It could be difficult even for a research officer to backtrack the employment history of a researcher.

The third stage, introduced in March 2010, denotes that a special librarian (one of the repository staff) does a complete bibliographic check and adds other relevant data to the publication. This includes information such as subject classifications, keywords and controlled vocabulary, and the full text if available.

Method

The method used to analyze the checking of accuracy of the metadata of repository items was to compare the history and note any changes of items on two separate occasions, before and after the implementation of the third stage in the repository workflow, the complete check by a special librarian. The changes to one or several individual fields of the item were analyzed to obtain data on frequent errors, missing data or missing full texts. This analysis provided valuable information for the repository staff's understanding of academic staff as users of the repository, as well as possible areas of improvement regarding the quality of the publication data. Total additions/changes/deletions for research publications from the year 2009 are 2052 for 698 items.

At the time of the implementation of the third stage - 2009 and the beginning of 2010 - Malmö University had not yet decided on an Open Access-policy. The pursuit of the full text was accordingly not in focus.

Results of the present study

During a large part of 2010, as mentioned earlier, a third stage in the acceptance procedures of MUEP was tested and evaluated. A report was published in June 2011[2] as an internal report. The subjects of the study were publications published in 2009. The aim and purpose of the project was to increase the quality of the registered publications in MUEP, by ways of adding, correcting and double checking the bibliographic information registered by researchers. The issue of quality is a central concern for an institutional repository in relation to national

[1] Lindholm, Jessica & Nilén, Peter: A New Open Access policy for Malmö University. ScieCom Info (2011) vol 7, no 1. http://www.sciecom.org/ojs/index.php/sciecominfo/article/view/4910

[2] Widmark, Jenny: Datakvalitet i MUEP. Rapport från datakvalitetsprojektet Bibliotek och IT2010. Revised August 2011 by Peter Nilén & Jessica Lindholm.(Unpublished report, 2011).Contact: jenny.widmark@mah.se

harvesting services as the Swedish SwePub[3] service, search engines like Google Scholar or Scopus, or as a base for bibliometric analyses of research publications (internal and external).

Another aim of the project was to establish a method or workflow that included a dialogue with the researchers in order to obtain a higher understanding of the importance in the quality of the registrations in the institutional repository. Malmö University is not research intensive, and we handle about 600-700 research publications a year.

The result of the project was the conclusion that a third stage in the acceptance procedures played an important role in increasing the quality of the bibliographic data/information. The third stage was then made permanent in January 2011, when a librarian with special focus on open access issues, as well as having the abilities to ensure the bibliographic quality of records in the system, was hired on 70% of a full time. A task takes between five minutes and two hours to complete, depending on the complexity.

This article is a continuation of the project outlined above as an analysis of the changes and additions made in the third stage for the publications in 2009. The aim has been to recognize the most common changes and additions and to use this information to design an approach or strategy for the repository staff in relation to the registering researcher.

The ten most frequent changes or additions to the items in the repository made in the third stage are (total additions/changes/deletions for publications from the year 2009 are 2052):

1. *Identifier citation added/change/removed: 17% (357 of 2052)*
   The citation information not correctly stated. Could also be the result of the researcher copying the citation information from a database. The string of information should be stated in different fields in the repository.

2. *Items added/removed: 12% (242 of 2052)*
   Publications added by the bibliometric department as a result of searches in external databases, i.e. publications not registered by researchers affiliated to the university. Removed publications could be items deleted if registered in the wrong collection in the repository. In 2009/10 it was not possible to change without deleting the item and re-registering it.

3. *Identifier ISSN added/change: 9% (190 of 2052)*
   ISSN information not stated by the researcher.

4. *Contributor author change/removed: 8% (171 of 2052)*
   Name format is not correct. Could be the result when the researcher copies the name format from a database, for instance PubMed lists authors with surname followed by space and initial letter in the name. In MUEP the whole name must be stated.

5. *Subject added/change: 6% (121 of 2052) and subject SRSC/VR change 4% (85 of 2052)*
   Both author generated subject terms or keywords and the official Swedish controlled vocabulary (Swedish Research Council) used in SwePub[4] were added. In 2009 the hierarchical drop-down menu implemented posed some difficulties for the researchers. They simply did not understand the structure, often choosing a general term instead of the more specific one, minimizing the potential for a more detailed search in external databases.

6. *Title change: 6% (119 of 2052)*
   Due to the researcher copying the title from a database, often a final dot is included. This is then removed in the third stage. This category could also include corrections of spelling mistakes.

7. *Identifier DOI added/change/removed: 4% (91 of 2052)*
   DOI-number missing or incorrectly stated.

8. *Description abstract added/change/removed: 4% (83 of 2052)*
   Often added to the category of peer-reviewed articles.

9. *Identifier URL added/change/removed: 3% (79 of 2052)*
   Often added to the category of conference proceedings.

10. *Publisher added/change/removed: 3% (78 of 2052)*
    Could be due to abbreviation of the publisher name. A majority of the changes appear in the category of book chapters and were the result

[3] http://swepub.kb.se/

[4] SwePub is a national service that harvests academic publications from institutional repositories at approximately thirty (October 2012) Swedish universities. http://swepub.kb.se/

of researchers misunderstanding the data entry form.

As mentioned earlier, Malmö University decided on an open access policy in November 2010. The policy did not stipulate retroactive registrations of full text for older registrations/items in the repository. In accordance the repository staff has not worked on the issue of full text documents to the material from 2009. Nevertheless we found it interesting to analyse if the full text material had continued to be added over the years. In November 2012 seven more full text documents had been added. Few researchers supplement their publications retroactively with the full text.

Conclusion

Analyzing the changes made to specific fields for items in our repositories presents us with valuable knowledge about the tricky parts of registering and the descriptions of publications made by our researchers. Repository staff is able to introduce changes to the data entry forms and to design outreach programmes or instructions to help researchers fully understand the bibliographic aspect of their publications and its relation to Google Scholar or national harvesting services.

The information about corrections to repository items also gives repository staff the opportunity to quality assure the registration process and together with the researcher and faculty/university research officers create conditions for a high level of metadata quality for the research publications of Malmö University. This includes a joint review of the quality of the metadata for the specific faculty. Revising the data entry forms and instructions in dialogue with the research organization is another field of improvement.

The result of this study of 698 publications from the year 2009 will hopefully make a small contribution to a more general approach to quality assurance of contents in institutional repositories at our universities.

Pablo Tapia Lagunas is Librarian at Malmö University, Library & IT since 1998. Works in the field of Digital Information Services and as liaison librarian to Urban Studies at the Faculty of Culture and Society. Alma mater: Lund University. Special interest: university library and IT infrastructure for research and education, innovation and networking as a method.