

DSpace Data Extraction Service for ‘Office’¹ Users

Jessica Lindholm and Robert Faling

This paper gives an overview of the background and development of EXTRAMUEP², a stand-alone web application for DSpace data extraction, an application that aims at motivating content-owners to deposit more of their works as well as facilitating dissemination and administration of research. The main feature of this tool is the conversion of DSpace data into ready-to-use files in the user’s ‘Office’ applications.

Introduction

Since 2005 Malmö University³ is using DSpace⁴ for parallel publishing of research publications and registration of student essays. The workflow for the majority of entries in the system relies on authors registering their works themselves, and members of faculty reviewing the records before they are accepted. The university’s installation of DSpace is named MUEP, Malmö University Electronic Publishing⁵. We have a steady flow of new items into the repository, as well as a growing need for internal services to DSpace, to ensure that data entered in DSpace is not only created and used, but more importantly re-used by the university itself.

Usage of the repository

As experienced by several institutional repository providers, researchers are generally hard to convince to deposit their content (e.g. Davis and Connolly 2007⁶). Malmö University is participating in a nationally funded project (SBF)⁷ during spring 2007, in order to find out more about how the researchers look at parallel publishing and workflows in institutional repositories.

Reasons for the non-usage of the repositories have been pointed out in several publications lately. Alma Swan, Key Perspectives Ltd, points out the main worries from the faculty staff that if they self-archive:

- they will be infringing a rights agreement with their publisher;
- it will take too much time; and
- it will be difficult to do

Swan also points out the researchers’ need to make an impact by communicating their results to their peers.⁸ The Swedish project is a response to papers mentioned above, with a specific focus on getting input on how to provide assistance to the researchers, as well as ensuring a copy in the local repository. In the in-depth interviews, we have received an explicit message from several researchers that, due to lack of time, parallel publishing in the institutional repository would only be prioritised if the organisation or funding agencies decided it to be

¹ The application is used and tested in e.g. Star Office, Open Office and Microsoft Office suites.

² <http://extramuep.mah.se/>

³ <http://www.mah.se/english/>

⁴ <http://www.dspace.org/>

⁵ <http://mah.se/muep/> (English and Swedish)

⁶ Davis M. Philip and Matthew J. L. Connolly (2007), “Institutional Repositories - Evaluating the Reasons for Non-use of Cornell University’s Installation of DSpace”, D-Lib Magazine, March/April, Volume 13 Number 3/4, <http://www.dlib.org/dlib/march07/davis/03davis.html>

⁷ BIBSAM-funded project, “Självarkivering och beslutsstöd för forskare vid publicering - en användningsfallstudie” (description only in Swedish), <http://www.kb.se/BIBSAM/bidrag/projbidr/pagaende.htm>

⁸ Swan, Alma (2006) “The culture of Open Access: researchers’ views and responses” in Open Access: Key Strategic, Technical and Economic Aspects. Jacobs, N. (ed.), Chandos Publishing, Oxford. ISBN 18-4334-2049

mandatory or strongly encouraged. As regards their needs for support in the task, ideally the paper was emailed to “someone else” who checked and controlled rights for parallel publishing, as well as depositing and describing the item. This is of course discouraging, but at the same time we need to recognise and accept the researchers’ needs to communicate research outside the organisation.

These responses made it clear to us that we needed to be pro-active and find other ways to ensure that research at Malmö University is made more visible and MUEP even more useful.

EXTRAMUEP - Facilitating Research Dissemination and Administration

Every year the university gathers information about the research conducted at the university. The information is used for distribution of research funding, as well as being gathered into a national compilation. The normal procedure for gathering this information is carried out via emails within the faculty, which results in a spreadsheet that afterwards is sent to the research coordinator at the university, where the data is used for statistical purposes and to determine keys for funding upcoming years.

The information in the spreadsheets is similar to the one we have in DSpace, and we could see the benefits of getting the data extracted automatically for this purpose, and in a way that does not require end-users to use other software than the ones they are used to in their desktop environment.

In late 2006, we allowed users to skip the file upload in DSpace, by activating a DSpace patch, “Configurable Item Submission” (available at <http://sourceforge.net>). This was necessary since the university realised that DSpace held data that could be used if the data was exhaustive. All bibliographical data on research publications in 2006 was added in DSpace by two large institutions at Malmö University, and after this library staff manually copied and pasted the data into spreadsheets, as an attempt to see if the metadata was good enough for the task. In the evaluations afterwards we found out that a majority of the researchers that had added their data in DSpace had a positive attitude to doing it again for 2007, and that the administrative staff could see a slight improvement of data quality⁹. We had implications to automate this task for coming years.

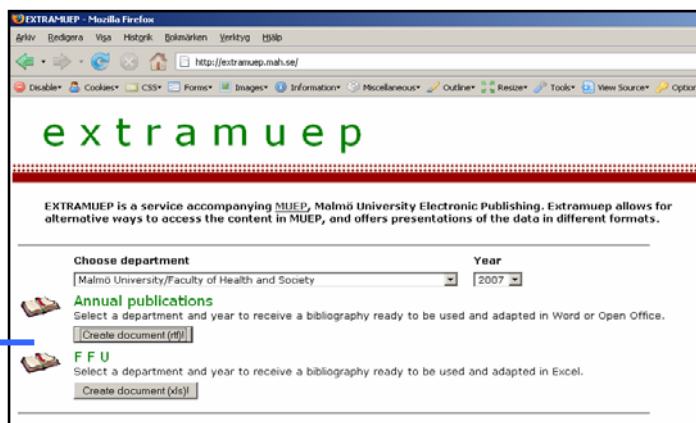


Figure 1. Screenshot from EXTRAMUEP <http://extramuep.mah.se>. From the drop-down menu the user chooses a year (mapped to dc.date.issued) and community (mapped to handle nr), and chooses an output format – currently (1) word processing file or spreadsheet.

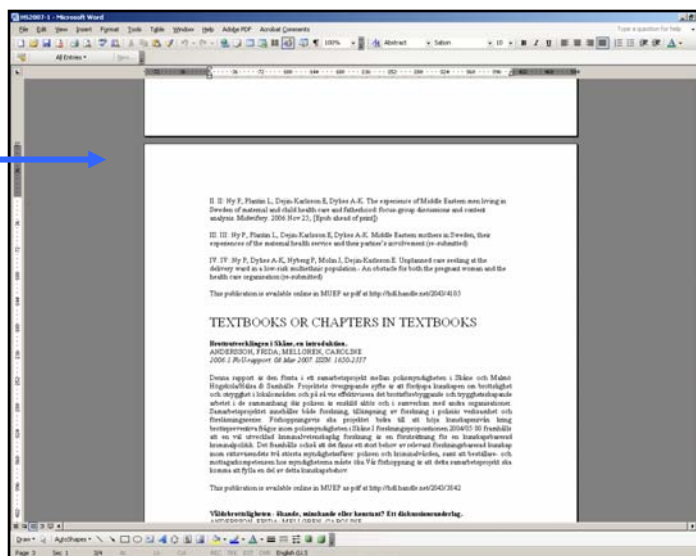


Figure 2. (2) Word file pop-up after request by EXTRAMUEP. The presentation is simply another way of viewing DSpace data, fields are familiar, we see titles, authors, abstracts, links to full-texts etc.

⁹ internal evaluation in Swedish - can be distributed by authors on request

Hence we developed EXTRAMUEP, a stand-alone application for extraction of data. For the end-users it is a web-based interface where the user downloads a file that is opened in their favourite software used for spreadsheets and word processing. The application itself will be able to be used for other formats, layouts and field extractions.

It is a stand-alone web-based PHP-solution, which calls the database directly and maps data to Rich Text Format and HTML 4.0 as the output formats. The application tells the browser to open the files in the software that is associated with the specific content types.

Conclusions and Further Development

Following the initially mentioned list by Alma Swan, with worries for researchers about self-archiving potentially infringing rights agreements, taking too much time as well as being difficult - we can see that EXTRAMUEP gives the benefit of saving time for internal registrations of data, and while adding data researchers realise that it is not too difficult⁹ to use.

We also find researchers that now regularly add their publications to DSpace. We have also seen an increased amount of full-texts being uploaded into the repository and data added that is not mandatory, such as abstracts and links to full-texts etc.

During summer 2007 we plan to develop the spreadsheet extraction service more, as well as a feature to extract HTML snippets to copy and paste into e.g. a researcher's web site, or other use. When EXTRAMUEP is finalised we will document and package the application for usage and further development by others in the DSpace community.

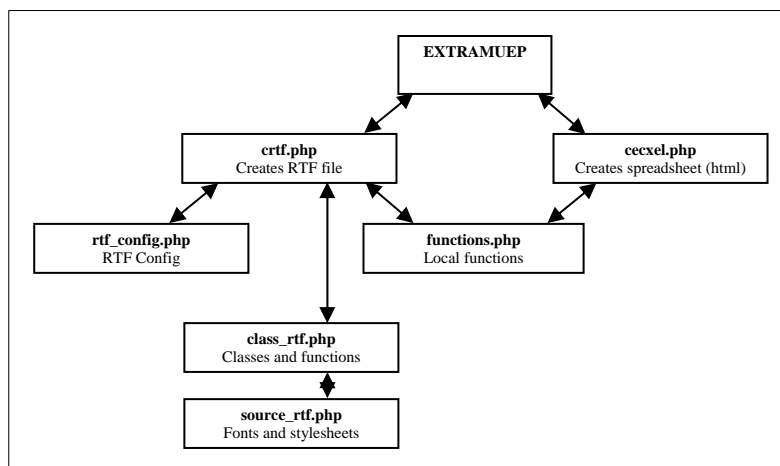


Figure 3. Current classes and functions in EXTRAMUEP

Autobiographical note

Jessica Lindholm, Master of Arts Library and Information Science, is a librarian at the Library and IT department at Malmö University. Jessica works mainly with coordination and development of e-publishing activities related to Institutional Repositories and Open Access. Jessica has been working with digital library development since 1999 in universities in Sweden, Denmark and in the U.K. Jessica's main interests are metadata functionality, information architecture and interoperability; in relation to the development of applications for digital libraries.

Robert Faling, Bachelor of Science in Electrical Engineering, is a developer at the Library and IT department at Malmö University. Among several other things Robert works with the technical development and administration of the DSpace installation at Malmö University. Robert started out at Malmö University as a student in Electrical Engineering 1996 and has worked at Malmö University the last couple of years with different assignments.