DASH for Collection Development Final Report

Tomoko Kurahashi and Reinhard Engels

Project Summary:
This project provides Harvard Institutional Repository, DASH, with the capability of a library collection development tool to aggregate reference data in PDF formatted articles deposited in DASH. These reference data were extracted from the articles into some bibliographic data segments, and these journal titles were compared with the Harvard Library collections in order to understand if the materials cited by Harvard faculty members are already held in Harvard Library. This comparison was reported with matching indicated or the lack of library holdings information over DASH for Collection Development project website and would support librarians to make a decision on Harvard Library collection. This was the first phase of the project, which focused on journal titles that are dominant resource in research work, especially in STM (scientific, technological, and medical) disciplines.

Accomplishments:

- Extracted reference data from PDF files in DASH using open source codes.
- Retrieved journal titles by parsing the extracted reference data.
- Created a list of the print and electronic journal titles plus journal abbreviations that Harvard Library subscribed or subscribing by utilizing Harvard Library Bibliographic Dataset (http://openmetadata.lib.harvard.edu/bibdata).
- Matched the extracted journal titles from reference data in DASH and a list of subscribed journal titles in Harvard Library holdings by running a script in Harvard's FAS supercomputer, Odyssey.
- Reported the matching results on DASH for Collection Development project website (http://osc.hul.harvard.edu/dash4coldev/).
- Linked the results of matched titles with Harvard Library collections and the number of their corresponding articles in DASH to its HOLLIS webpage and DASH webpage respectively in order to quickly access the original resource.

Challenges:
The most challenging work was to obtain properly extracted journal titles from the PDF files in DASH. There were several conditions that seems to corrupt the data. Firstly, the papers in a particular discipline use a particular reference citation style. Each citation style is slightly different from the other. Secondly, some reference sections are placed at the end of the papers, but some are as footnotes throughout the papers. Thirdly, some papers mix the reference and annotation together. Fourthly, some reference citation were unfortunately written wrongly or mis-typed. Finally, abbreviations of journal titles are heavily used in some particular disciplines. The above issues affected PDF extraction and consequent processes.

Next Steps:
Because of imperfect data from PDF extraction, it is necessary to improve the accuracy of the PDF extraction before moving on. It could be useful to explore the extraction of reference data in the PDF articles with DOI links or other data segments pointing to external systems, which may help to retrieve more accurate bibliographic data.

Some subscribed journals skip parts of the chronological periodicals, and some are also provided differently by different vendors simultaneously. Therefore, the publication years or volume numbers of each journal title from the reference section would need to be considered for more detailed comparison between references in the papers in DASH and Harvard Library holdings.

The second phase would focus on books and conference papers, which are the next dominant research resources led by journal articles in STM fields. And then, monthly report would be automatically generated and sent to the librarians who register for it. Visualization of the results also needs to be created in order to help librarians to understand and evaluate their collections effectively. In order to reduce the manual scripting steps, it would be good to develop automatic PDF extraction during each deposit or via monthly deposits.

Finally, this project has a potential to develop several applications, such as inter-disciplinary research and changes of the subjects through times, by analyzing subjects, publication years, and other related data in the papers in DASH and external resources. These developments will add more functions and value to DASH and other institutional repositories.

Budget Spent:
Total development hours: 80.5
Total development cost (including wage, technology use, etc): $

Publicity:
Demonstrated at the Library Lab Showcase on November 15th, 2012
Presented at the Library Lab Project Lightning Talk on July 27th, 2012