Harvard's DASH for Collection Development

Changes made to the first draft:
- Added Section 10: Conversations with Mr. Hazen (Associate Librarian of Harvard College for Collection Development) and Section 11: Applications
- Modified Section 5 in order to link to Section 11.

1. Author

Tomoko Kurahashi (Library Technician at Wolbach Library)

2. Background

Library collections including e-resources are assessed by several library usage statistics, such as circulation, in-house use, InterLibrary Loan (ILL), document delivery, COUNTER, and so forth. But, it reveals little information about how and what materials researchers exactly use them for their research.

On the other hand, citation analysis, analyzing what articles are most cited, such as Impact Factor (IF) and H-Index, are widely used to measure the importance or popularity as well as productivity and impact of the research (Wikipedia, n.d.). However, generally these analyses are provided by bibliographic databases, such as Web of Science and Scopus, which contains selected journals and proceedings, but not the journals, conference papers, or lecture notes that are not indexed in these databases.

Since 2009, Harvard's faculty have been depositing their research publications into the Harvard Institutional Repository, DASH (Harvard University Library, 2009). Each publication contains a list of references that researcher actually used in the process of their studies. Therefore, if the references data of these publications can be stored in DASH and be compared to the Harvard library collections, librarians can figure out the gap in the collections and create some statistics and analyses to make a purchase/assessment decision for collection development.

3. Project

This project is to give DASH the capability of a library collection development tool. By aggregating references data from the articles deposited in DASH and comparing the Harvard library collections, the following questions could be answered:

- What journal/book titles are used?
- What journal/book titles do the Harvard Libraries lack in their collections?
- What are the core journals for a particular research/professor/department?
- What subjects are heavily used or rarely used?
- What subjects(LC numbers) are the weaknesses/strengths of a local library?

This project has 2 development phases.

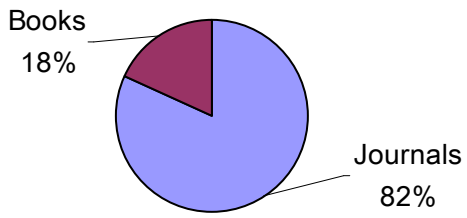
Phase 1:
Extract journal titles, which are most used resources, from PDF files in DASH, check against HOLLIS journal titles by some scripts, and generate a report.

Phase 2:
Like Phase 1, extract titles from non-journal resources as well as keywords in the articles from PDF files in DASH, compare to HOLLIS records, and generate a report (see example).

Report Example

Wolbach Library; The materials used by FAS Department "Astronomy" (09/2011-10/2011)

Material types:



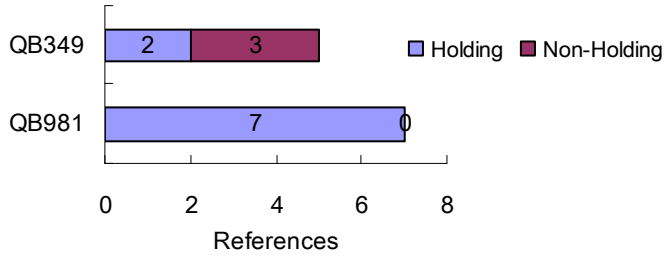
Journals:

Journal Title	References in DASH	Harvard Current Subscription
Journal of A	36	Y
Journal of B	13	Y
Journal of C	3	N
Journal of D	1	Y

Books:

Book Title	Call No.	References in DASH	Harvard Holding	Local Holding
E book	QB981	7	Y	Y
F book	QB349	3	N	N
G book	QB349	2	Y	Y

Call Numbers (Books):



This report could be shown on the site or be sent to librarians as a report.

4. Purpose

The purpose of this project is to support Harvard library community to have more accurate data of faculty-used materials, to help the community make decisions for collection development, and to increase the functions of DASH.

5. Benefits

By applying DASH as a collection development tool, librarians could gain more concrete information about what materials researchers use exactly in their research and find the gap between these materials and library collections. Also, this project uses existing productions, such as articles created/deposited by Harvard community, and does not create a new production or new theory so that results of the analysis will be more relevant and reliable to the Harvard community.

In addition, DASH will automatically compare the references data to HOLLIS records, and hence it will reduce compilation time for librarians to assess the library collections. More enhanced library collections will enrich Harvard community's scholarly activities. Furthermore, this project will extract data from PDF files in DASH so that other applications that use these data could also be easily evoked. More detailed information about these applications can be found in Section 11.

6. Resources

Although additional research is needed, the following resources/technology can be used for this project:

- DASH
- HOLLIS data
- Tools for extracting references from a PDF file:
  http://code.google.com/p/pdftoref/
  https://code.google.com/p/pdfssa4met/
- Tools for comparing bibliography/references data and HOLLIS data
- Methods to interpret the subject(s) of each citation, specifically journals

7. Schedule

Phase 0: Research (1st month)
Phase 1: Journals (2nd-3rd month)
Phase 2: Non-journals and other data (4th-6th month)

8. Budget

1 Project Assistant: 100 hours = $3000
1 Digital Library Software Engineer: 200hours/Release time

9. Measurements for Success

- Less time for collection development librarians to analyze the collections
- Positive feedback from Harvard library community
- The number of other universities libraries that follow and modify their institutional repositories (IRs) from this project idea

10. Conversation with Mr. Hazen at Widener Library

There was an opportunity to meet Mr. Hazen, the Associate Librarian of Harvard College for Collection Development, for this project. He mentioned several pointers to this project, such as how open access journals, different classification in the libraries (e.g. Old Widener System at Widener Library) etc need to be considered. But he also expresses his strong interests in this project that it can analyze how researches are inter-disciplinarily related and how the subjects are changing over time. He believes this project will be the first step for these analyses and supports it.

11. Applications

There are also several applications as a result from some other meetings/communications.

- Journal characteristics
- Inter-disciplinarity of research
- Changes of the subjects of the times
- Identifying the citations by Harvard authors
- Applying to HMScholar
- Finding prospective collaborators

These are more advanced and, in some cases, need more time to invest on the methodologies, such as how to map the Harvard authors to the authors in references in the articles, etc. Therefore, this project will primarily focus on more critical, essential information for collection development needs, such as journal titles in references, and will explore the above work in phases.

12. Conclusion

The collections used by researchers should be clear to collection development librarians. This project not only provides them with more powerful tools, but also reduces their workload. IRs for library collection development has not been practiced by any institution yet. As the Harvard library community, this project will be the first implementation of collaboration between IR and collection development and will stimulate DASH and the entire Harvard community.

References

Harvard University Library. (2009). Harvard's DASH for Open Access. Retrieved from http://hul.harvard.edu/news/2009_0901.html

Wikipedia. (n.d.). h-index. Retrieved from http://en.wikipedia.org/wiki/H-index