Access Flows to a Repository from Other Services

Daisuke Ikeda
Department of Informatics, Kyushu University
daisuke@i.kyushu-u.ac.jp

Sozo Inoue
Kyushu University Library
sozo@lib.kyushu-u.ac.jp

Abstract

In this paper, we analyze access logs of an institutional repository and show access flows to the repository. In addition to the log files of the repository, we also use those of the *linking system* which connects the repository and the activity database of faculty members in the institution. We show that many accesses come from search engines, such as Google, and only few accesses from harvesters, such as OAIster. This fact leads us the tentative hypothesis that metadata of repositories does not work for searching items on them. In spite of that, this fact implies external services play a important role to let users reach to repositories. In fact, we reveal potential importance of external services with large page views. Such services play as a role of pathfinders to items on repositories.

1. Introduction

Repository statistics by OpenDOAR [1] show world-wide deployment of institutional repositories: more than 1,300 repositories including 72 in Japan are deployed as of 20 Feb. 2009. In Japan, we find 92 repositories in “Current IRs”[2]. From these figures, we can conclude that the institutional repository has become widespread.

As the institutional repository has become popular as described above, evaluation of repositories is becoming more important. Many attentions have been paid on input status of a repository, such as the number of contents on the repository [1]. The output status, such as access analysis, is becoming much more important as repositories are compiling many items. In this paper, we analyze access logs of our repository and show access flows to the repository.

In addition to log files of the repository, we also use those of the *linking system* which connects the repository and the activity database of faculty members in Kyushu University. The linking system links from a paper [3] in the activity database to the corresponding item on our repository and conversely it links from an author of the repository to his/her page of the database. The faculty members are obliged to input their activity data into the database and so it boasts high data coverage and large page views.

Although access logs of a repository only show explicit demands to academic works which have already contained on the repository, our analysis with access logs of the linking system can unveil potential demands to academic works which are not on the repository. In fact, our analysis will show that our repository does only contain papers for about 13% of requests.

In the literature, many efforts of access analysis have been made for electric journals [9, 10] or inter library loans [13], because huge number of logs for these services have already compiled. On the other hand, access log files are now compiling for institutional repositories [8, 14], and so we do not have systematic researches on the access analysis for repositories.

Basic findings by our access analysis are that more than 50% of accesses come from search engines but only few accesses from harvesters, such as OAIster. We can say that contents of items on a repository is much more important than metadata of them from the viewpoint of searching, although it is originally thought that harvesters and the metadata exchange protocol OAI-PMH play important role for searching institutional repositories.

We also show that other paths to a repository: there exist stable accesses from the activity database and CiNii, which is a portal of academic papers in Japan, to our repository. These facts lead us the tentative hypothesis of division roles: search engines take a major role for searching an item on a repository directory and external services plays a role of pathfinders to a repository. We can expect many external services will be linked with repositories [5,6,12]

[1] http://www.opendoar.org/

[2] http://www.nii.ac.jp/irp/list

[3] The activity database only contains metadata of the paper.

and hence this hypothesis will be verified in this new environment.

2. Linking System

Academic Staff Educational and Research Activities Database [4] records activities of faculty members in Kyushu University. As a part of the database, it compiles lists of papers. Since faculty members are obliged to input their activity data into the database, the number of papers of the database is much larger than the number of items on our repository QIR[5] which stands for Kyushu University Institutional Repository.

The activity database has huge page views but it compiles only the metadata of papers. The linking system provides a link from a paper in the activity database to the corresponding item on our repository and conversely it links from an author of the repository to his/her page of the database [11]. As one of the important features of the linking system, it provides a link even if the corresponding item does not exist on the repository. In this case, if a user click the link, then he/she is moved to the registration process of the paper to the repository. Thus, we can estimate potential demands for academic work which are requested from the activity database but not on QIR.

We developed “Footprint” Visualization system which shows a faculty member the number of clicks for unregistered papers of the member [7].

3. Usage Analysis

In this section, we analyze the accesses of our repository called QIR.

The period of the log files we will use is from 2008/07/01 to 2009/01/31. We preprocess log files to eliminate accesses from crawlers and count if an access has been successfully done, that is, if an access has “200” or “304” status code of the http protocol.

3.1. Access from Search Engines

Firstly, we count the number of accesses from search engines.

Figure 1 is the graph of the number of accesses: red bar shows the total number of accesses to QIR and blue one the number of accesses to QIR via search engines.

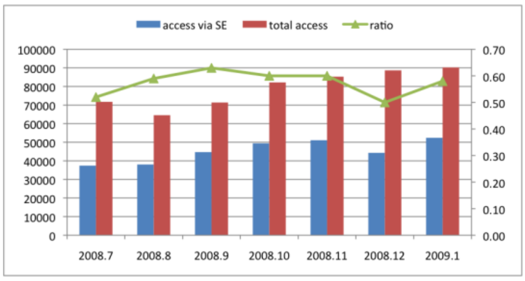

Figure 1. The graph shows access figures of QIR and those of QIR from search engines and their ratio. The left axis shows the number of access and the right one the ratio.

We find that more than half of total accesses come from search engines.

3.2. Access from Activity Database

Next, we focus on accesses from the activity database. We use the log files of the linking system, which records accesses between the activity database and QIR. The linking system put hyperlinks of papers on the database, with user’s permission, even if these papers are not on QIR.

Figure 2 shows the accesses of QIR from the database.

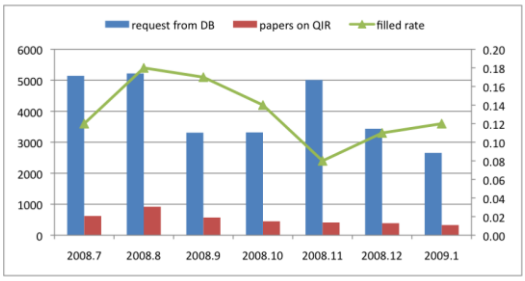

Figure 2. The graph shows the number of clicks on links from the activity databases to QIR (blue bar), the number of successful access to papers among these clicks (red bar) and their ratio (line chart)

The blue bar in the graph shows the number of clicks at the database and the red one the number of accesses from the database to QIR. This means that QIR only provides about 13% of papers against all requests from users of the activity database.

From this graph, we can also find that averagely more than 4,000 accesses per a month come from the activity database.

[4] http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html

[5] http://qir.kyushu-u.ac.jp/

3. Access from Harvesters

Next, we focus on accesses of QIR from harvesters, such as OAIster. We consider the following 6 sites as harversters, where CiNii is not a pure harvester but has been a portal of academic papers in Japan.

Scirus http://www.scirus.com/
OAIster http://www.oaister.org/
ScientificCommons
http://www.scientificcommons.org
Jairo http://jairo.nii.ac.jp/
JuNii+ http://ju.nii.ac.jp/
CiNii http://ci.nii.ac.jp/

It compiles papers published in academic society journals or university research bulletins, or included in the National Diet Library's Japanese Periodicals Index Database. Therefore, it has had huge page views before it started to provide a service of the harvester since 2008/10/06.

Figure 3 is the graph of accesses of QIR from the activity database and above harvesters, where "DB(list)" stands for accesses from links on papers of the database and "DB(top)" stands for accesses via QIR's RSS gadget on the database interface. Thus, the sum of these two accesses is the total number of accesses to QIR via the activity database. The average of access numbers via the database is around 4,400.

The numbers of accesses via harverster from July to September 2008 are quite low, monthly average is around 300. But, after October these accesses have sharply increased. This is due to CiNii, which originally has had the ability to pull in many users.

4. Conclusion

We have analyzed log files of the linking system between our repository and activity database, in addition to log files of the repository.

We have found that many accesses, more than 50% of the total accesses to items on the repository, come from major search engines, and the small number of accesses come from harvesters. Our analysis has also showed that other services related to academic papers or activities provide strong and stable accesses to our repository via these services.

We have showed that many users do not use metadata when they search academic work on institutional repositories. However, this fact does not immediately lead to the conclusion that metadata is useless, because metadata demonstrates its ability when many services including repositories works together cooperatively [2,3,5] and items moves these services automatically [6,12].

Acknowledgments

This research was partially supported by the National Institute of Informatics, Institutional Repositories Program 2008 [4] and the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300082, 2008.

References

[1] Leslie Carr and Tim Brody. Size Isn't Everything: Sustainable Repositories as Evidenced by Sustainable Deposit Profiles. In *D-Lib Magazine*, volume 13, July/August 2007.

[2] James Dalziel. Access Management: Challenges and Approaches. December 2003. http://www.library.usyd.edu.au/dest/dalziel.ppt.

[3] James Dalziel. Integrating Identity Management - Aspirations and Issues. 2006. http://www.apsr.edu.au/Open_Repositories_2006/james_dalziel_2.ppt.

[4] Daisuke Ikeda and Sozo Inoue. A New, Sustainable Model for the Institutional Repository: A CSI Project "Integration and Presentation of Diverse Information Resources". In *DRF International Conference Open Access and Institutional Repository in Asia-Pacific*, February 2008.

[5] Daisuke Ikeda and Sozo Inoue. A Sustainable Model based on the Social Network Service to Support the Research Cycle. In *Proceedings of The 3rd International Conference on Open Repositories*, April 2008.

[6] Daisuke Ikeda, Takashi So, Sadayoshi Noutomi, and Sozo Inoue. The Versioning Facility of Institutional Repositories as Support Tools for Research Activities. In *Digital Library*, volume 33, pages 31–38, November 2007. (in Japanese).

[7] Sozo Inoue, Tatsuro Fujii, Ken Kozai, and Daisuke Ikeda. Footprint Visualization for Motivating Academic Infomation Repositories. *Kyushu University Library, Research and Development Division Annual Report*, 2007/2008:17–22, 10 2008. (in Japanese).

[8] Christine Merk and Nils K. Windisch. Jisc usage statistics review: Final report. Project

Report, 9 2008.

[9] David Nicholas, Paul Huntington, Hamid R. Jamali, and Anthony Watkinson. The Information Seeking Behaviour of The Users of Digital Scholarly Journals. *Information Processing & Management*, 42(5):1345–1365, 2006.

[10] David Nicholas, Paul Huntington, and Anthony Watkinson. Scholarly Journal Usage: The Results of Deep Log Analysis. *Journal of Documentation*, 611(2):248–280, 2005.

[11] Mayumi Ono, Sozo Inoue, and Nami Hoshiko. Linking Institutional Repository and Researchers Database in Kyushu University. In *Digital Library*, 11 2007. (in Japanese).

[12] Matthias Razum, Frank Schwichtenberg, and Rozita Fridman. Versioning of Digital Objects in a Fedora-based Repository. In *German e-Science Conference*, May 2007.

[13] Syun Tutiya, Hiroya Takeuchi, Yoshinori Sato, and Hiroshi Itsumura. ILL/DD in Japan Across the Turn of The Century: Basic Findings about NACSIS-ILL from 1994 to 2005. *Progress in Informatics*, 4:1–21, 2007.

[14] Chiba University. Standardization of Usage Statistics for IR Evaluation. http://www.ll.chiba-u.ac.jp/~joho/CSI/standardization.html, 2008. (in Japanese).

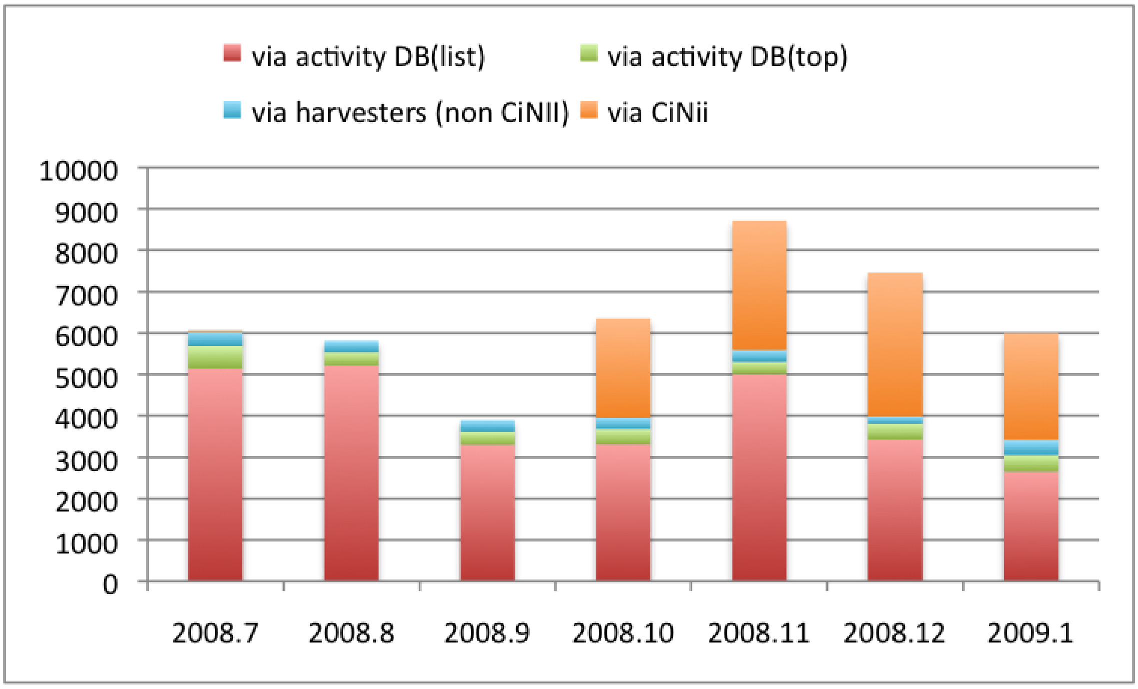

Figure 3. Accesses from the activity database and harvesters. "DB(list)" stands for accesses from links on papers of the database, while "DB(top)" stands for access from links from QIR's RSS gadget on the database interface.