Vol.2011-MPS-85 No.29
2011/9/16

Cooperation between Repository and Researcher Database

Kensuke Baba,[†1] Toshie Tanaka,[†1] Emi Ishita,[†1]
Masao Mori,[†1] Eisuke Ito[†1] and Sachio Hirokawa[†1]

This paper introduces a Web system which activates institutional repositories and evaluates the effect of the system. Institutional repository is an important service of libraries in academic institutions. The authors have developed a link system between the institutional repository and the researcher database of their university. By the developed system, some papers in the repository are linked from the metadata of the papers in the researcher database. Additionally, the system reuses the metadata of papers registered in the researcher database for paper registration to the repository. The authors also developed a function which analyzes the access log to the repository and returns the result to the authors as a feedback from readers by using the link system. The authors observed the log of download and upload on the repository before and after the start up of the system. The result shows that the system increased the number of access but there was no significant change of the number of registration.

1. Introduction

"Open access[14)]" to scholarly information provides free availability of research outputs such as scholarly papers. Generally, it seems to be reasonable the opinion that the research outputs funded by public institutions should be returned to society. Actually, the National Institutes of Health (NIH) showed their policy which requires the researchers funded by NIH to open their research outputs[6)]. One of the methods to realize the idea of open access is "self archiving"[13)]. Then, a *repository* is a system to archive and open research outputs, and a repository for outputs in an institution is called an *institutional repository* (*IR*). By improving the IR in each institution, open access to scholarly information can be realized.

A problem of IR is the fact that the number of the archived papers is extremely small compared to the papers practically produced by researchers. For example, the ratio in the IR of Kyushu University (QIR)[3)] is less than 30 %[10)]. Since the number of the papers in the IR ranks 51st in about 2,000 institutions in 4) as of July 2011, most institutions are considered to be in the same situation. The first step to improve IR is to encourage researchers to register their buried papers and prevent burying current papers. Although several kinds of trial are being made to increase the number of papers in each IR, it is difficult to apply the solutions to other IRs. Therefore, it is significant for any IR to describe formally a practical system which improves an IR.

If researchers are forced to register their papers to IR as a mandate, the problem of the number of archived papers may be solved. However, every institution cannot apply this solution immediately. An approach to the problem is to reduce the efforts of authors (that is, researchers) for paper registration to IR. One of the straightforward solutions is to archive only the metadata (that is, information about the title, the author(s), and so on) of a paper and URI of the full-text in an external Web site. However, this solution requires an agreement of a subscription with the site of the full-text, which does not realize the situation of the basic idea of open access. Another approach is to make an incentive for researchers to register their papers to IR. A simple incentive is an increase of the number of access to their papers. To urge researchers to register their papers to IR, more returns are necessary.

In this paper, we organize the improvements made for the IR in Kyushu University. As a solution based on the first approach to the problem of the number of archived papers, we developed a system to reduce the efforts of authors for paper registration to the IR by connecting with the researcher database (DHJS[⋆1])[2)] in the university[9)]. By the system, the metadata of a paper already registered in DHJS is reused for registration of the paper to QIR. As for the second approach, the system links the metadata of papers in DHJS to the full-text in QIR, which increases the number of access to the IR[12)]. As another solution based on the second approach, we are analyzing the access log of QIR[8)] and developing a system to feedback the result to authors[11)]. We also evaluate the effectiveness of the improvements by analyzing the access log to the IR in terms of the number

†1 Kyushu University

⋆1 "Daigaku Hyoka Joho System" in Japanese

of access and the number of registration.

The number of the institutions who have own repository in the world is about 2,300 as of August 2011[5], and most of the institutions are considered to have the same problem. In this paper, the situation of the practical systems in Kyushu University are shown in detail, and the problem and solution are described formally. Therefore, the main idea of the proposed system is applicable to other institutions.

2. Problem

This section describes the basic information of two databases in Kyushu University, QIR and DHJS, to make clear the problems we tackle.

2.1 QIR

QIR is the IR based on DSpace[1] and operated by Kyushu University Library. Generally, IR archives the full-text of each paper in addition to its metadata. Figure 1 is an example of the Web interface of the IR. In the interface, the word "∗.pdf" is linked to the full-text of the paper and the name of each author is linked to the profile page of the author. The total number of the papers in QIR is about 17,000 as of July 2011. Ranking Web of World Repositories is taking account of the number of the full-text files as an element of the ranking, then the number of QIR ranks 51st as of July 2011. Since the scope of the ranking is about 2,000 IRs, in most of the IRs the number of the archived papers are less than the number.

There exist two ways to register a paper to QIR:

- Create an account of the IR and register by the registration form,
- Submit the metadata and the full-text to the managers of the IR by Email.

In the registration form in QIR, the author has to fill the metadata and upload the full-text of their paper. Usually the operation to write the metadata of papers should be repeated when researchers upload the papers on their web-site, submit a list of the papers as a report to their institute, and so on, which may be an obstacle to paper registration to IR.

2.2 DHJS

DHJS is the researcher database of Kyushu University. The database has various kinds of data of the researchers in the university, for example, the posts, their

Fig. 1 An example of the Web image of QIR.

research interests, and the scholarly papers they produced. The number of the researchers in the university is about 3,000 as of October 2010. DHJS consists of the two subsystems, the data-entry system and the viewer system. The data-entry system supports researchers to register their research activities to DHJS and equips a user (that is, a researcher) identification by a password. The viewer system shows the research activities registered in DHJS by the data-entry system. The registered data is separated with respect to each researcher and the research papers of a researcher are described as a list in the viewer system. Fig. 2 is an example of the list of the metadata of scholarly papers shown on DHJS. The icons in the figure are mentioned in the following section.

In Kyushu University, any researcher has a duty to register their research activities includes the metadata of published scholarly papers into DHJS. Therefore, the database has the metadata of most papers produced in the university. The number of the "metadata" of scholarly papers registered in DHJS is about 86,000 as of July 2011. The ratio of duplicate data (that is, metadata for the same paper) is estimated at most 20%[10], hence the number of distinct papers is about

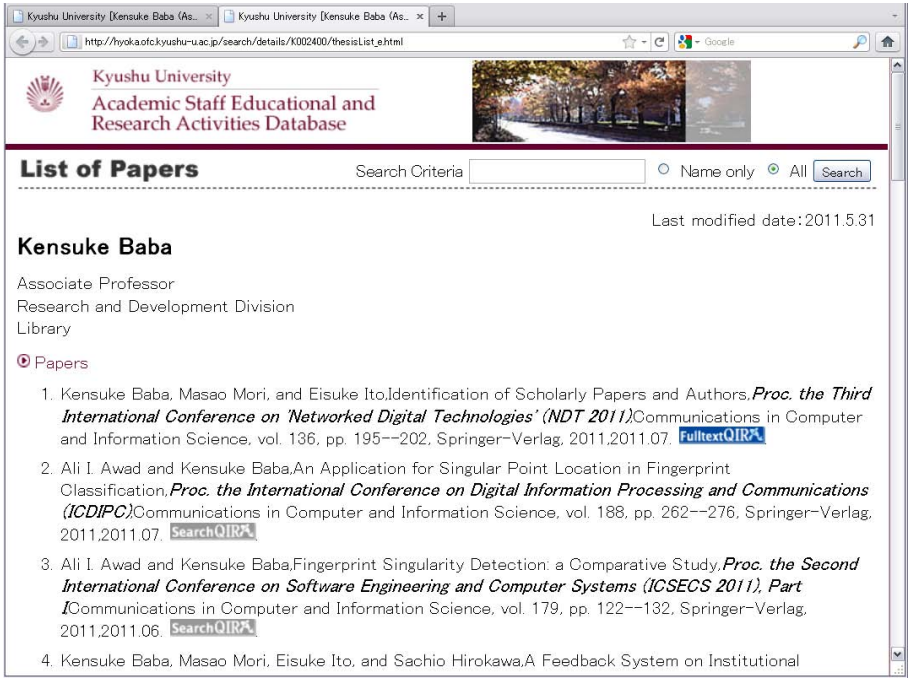

Fig. 2 An example of the list of scholarly papers in DHJS.

86,000. On the other hand, QIR has only 17,000 "full-texts" as mentioned in the previous subsection. That is, potentially, there exists a large number of research outputs which are produced in Kyushu University but are not archived in QIR.

3. Improvements

This section introduces two improvements made for QIR by connecting with DHJS. We considered two approaches to encourage researchers to register their papers to IR:

- To reduce the efforts of researchers for paper registration to IR,
- To make an incentive for researchers to register their papers to IR.

3.1 Link from DHJS to QIR

We developed a system which connects QIR with DHJS[9]. Figure 3 is the

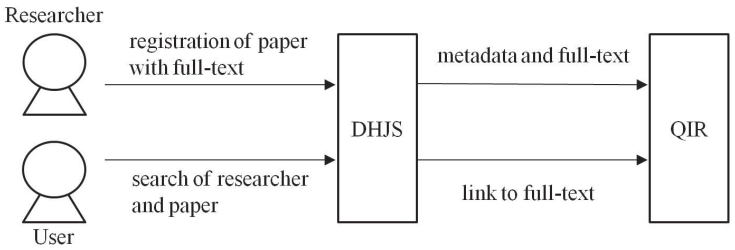

Fig. 3 The outline of the link system between QIR and DHJS.

outline of the developed system. The system realizes two functions:

- Paper registration to QIR is made in the data-entry system of DHJS,
- The metadata of papers in the viewer system of DHJS are linked to the full-text in QIR.

The first function is a solution based on the first approach. In the data-entry system of DHJS researchers can upload the full-text in addition to the metadata of their paper, which can reduce the efforts of researchers to register their paper to the IR.

In the data-entry system, researchers can make also a link from the metadata of a paper in DHJS to the full-text of the paper in QIR. The link is realized by an icon in the viewer system of DHJS. There exist two kinds of icons (see Figure 2) which distinguishes whether the paper is registered in QIR. For each paper in the list, there can be three kinds of situations,

- A dark-colored icon "FulltextQIR" is added,
- A light-colored icon "SearchQIR" is added, or
- There is no icon.

The first case means that there exists the full-text of the paper in QIR, the second that there is no full-text (although the researcher wants to register), and the other that the researcher does not want to link to full-text. In the first case, the user of DHJS can obtain the full-text corresponds to the metadata from QIR, which is expected to increase the number of access to the full-text in QIR. A large number of access to papers can be an incentive for researchers to register their papers to IR, that is, the function is a solution based on the second approach. In the

second case, the link system returns the result of a search by the author name in QIR. Additionally, in the case where the user is an author of the paper, the link leads the user to the registration form of QIR. At the time the metadata of the paper is automatically used to fill the registration form, therefore this function reduces the efforts for registration, that is, a solution for the first approach.

3.2 Feedback to Authors

Our solution based on the second approach is to analyze the access log of the IR and return the result to researchers as a feedback from the readers of their research outputs[11]. Then, the researchers can obtain the knowledge of reader's interests, which is instructive for spotting a research trend.

Some basic analyses of access log can be applied by DSpace, Google Analytics, and so on. For example, we can count the total number of the access for each paper and show the ranking on the IR by some basic functions on DSpace. Google Analytics can collect statistics about the region of the referrers of access, and the keywords if the access comes from the result of a search engine. In addition to the basic analyses, we focused on co-occurrence of access[8]. Figure 4 is an example of the Web image which shows the result of the basic analyses. The graph describes the number of the access to the items of the user and the top 10 user in the university. The horizontal axis shows the months and the vertical axis the number of the access. The table is the ranks of the number in the department of the user and in the university for each month.

A problem of implementation of the feedback system is that some analyses related to the authors make a kind of individual information. (Note that this problem is different from one for individual data of reader which can be obtained from the access log such as the IP-address.) For example, as to the ranking of the access and the keywords at the referrers for each researcher, some researchers do not want be open. Especially for the ranking, some researchers are worrying that the ranking would be used for assessment of the researchers, rather than the typical privacy problem. Actually, the simple total of the access in IR may not be suitable as a criterion for papers or researchers at present, although there seems to be a correlation between the number of access and the number of citations.

To solve the problem, the access to the result of the analyses should be controlled. The system we are developing utilizes the identification function of DHJS.

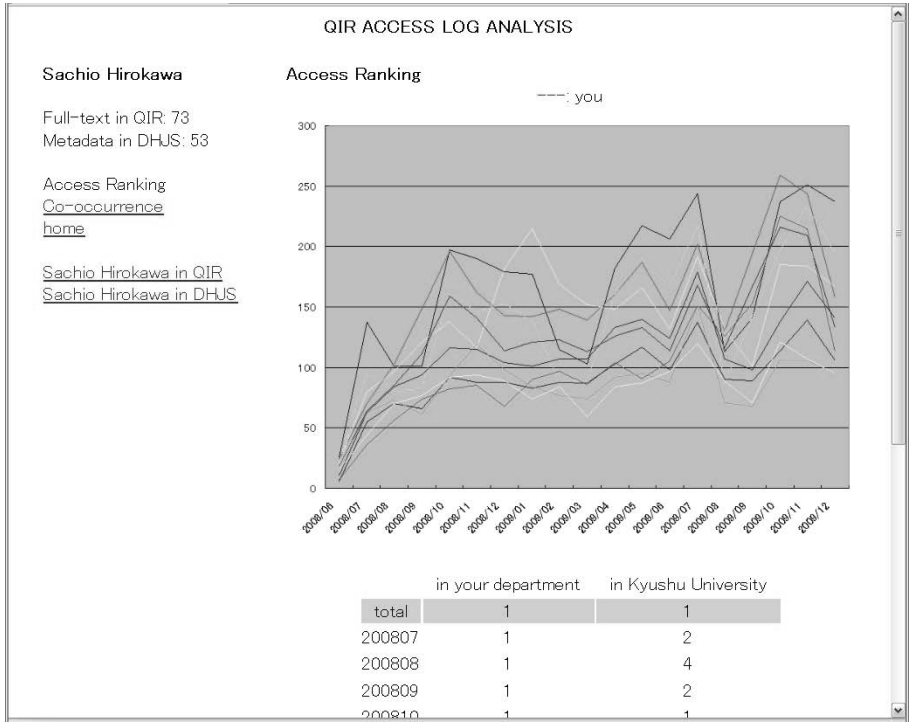

Fig. 4 The result of the total number of the access to the items of an author and the ranking.

Although QIR also has an identification function of users, the number of the users who have the account of QIR is small. On the other hand, the registration to DHJS is a duty of any researcher in the university as we mentioned. The Web image in Figure 4 is shown for a particular user only.

4. Evaluation

We evaluate the effect of the developed system in the two points, the number of access and the number of paper registration, by analyzing the practical access log. The number of the links which has the full-text (that is, the number of the colored icons) is 597 as of October 2010.

We analyzed the access log from June 2008 to October 2010. The total number

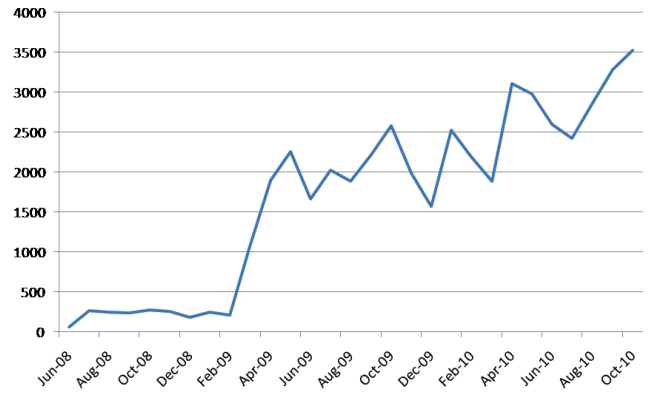

Fig. 5 The number of access to QIR via DHJS.

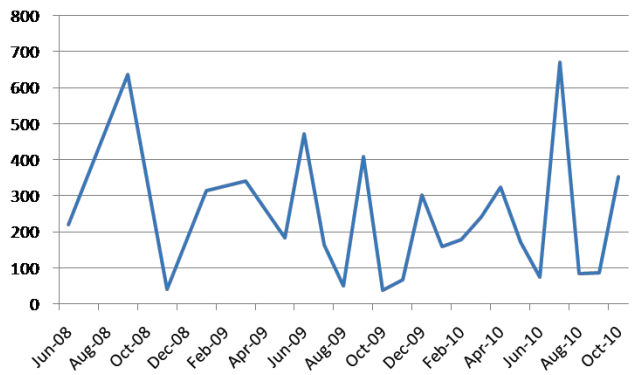

Fig. 6 The number of registered items to QIR.

of access is 40,975,341 and the average for a month is about 1,500,000. Figure 5 shows the number of access from DHJS to QIR for each month. The link system in Subsection 3.1 is working from April 2009 and was improved in April 2010. There are two points of increases, one is at March 2009 and the other is at April 2010. The ratio of this numbers to the total number of access is at most 0.21 %. Even after deleting the access of bots, the ratio is at most 0.34 %. This ratio is extremely small. However, since the total number of the links is 597, the amount of the increase (about 3,000 per month) is significant for the 597 papers. By the result of Figure 5, we can conclude that the link from DHJS is effective to the number of access to QIR. Although the ratio to the total access number is small, an obvious effect is expected by increasing the number of links.

We also analyzed the log of registration for QIR. Since the feedback system in Subsection 3.2 has not been operated yet as of October 201, the evaluation is only for the link system. Figure 6 shows the number of registration of papers to QIR for each month. As we mentioned the number of the links from DHJS is 597, therefore the number of registration by the link system is at most 597. Since the increase of the number from June 2008 to October 2010 is 7,411, the effect by the link system is estimated to be small. Actually, there is no significant change at April 2009 nor April 2010 in Figure 6. As the result, we could not find any effect by the link system in the sense of the number of registration.

5. Conclusion

The cooperative system between the IR and the researcher database in Kyushu University was introduced as a case study. The system was evaluated by analyzing the access log to the IR in terms of the number of access and the number of paper registration. As the result, we found that the system is effective to the number of access. However, we could not find any significant effect for the number of registration.

The previous result seems to show that reducing the efforts of paper registration is not so effective to increase the number of registration. Basically the paper registration to IR is made by researchers voluntary, however some researchers who are not interested in open access or IR may not register their paper even if the efforts of the registration be reduced drastically. It is necessary to inform widely about IR to researchers as a possible choice in addition to reducing the efforts, although it is difficult to obtain the consent to the idea of open access from every researchers.

Another problem to increase the number of paper registration to IR is the fact that the number of registration practically depends on the progress of handwork by repository managers, such as, confirmation of the copyright policy for each

Vol.2011-MPS-85 No.29
2011/9/16

paper. This can be a bottleneck even if a lot of requests of paper registration are made from researchers. We are developing a system to automate some processes by repository managers[7].

References

1) DSpace. http://www.dspace.org/, [accessed 19 Aug, 2011].
2) Kyushu University Academic Staff Educational and Research Activities Database. http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html, [accessed 19 Aug, 2011].
3) QIR: Kyushu University Institutional Repository. https://qir.kyushu-u.ac.jp/dspace/, [accessed 19 Aug, 2011].
4) Ranking Web of World Repositories. http://repositories.webometrics.info/, [accessed 19 Aug, 2011].
5) ROAR: Registry of Open Access Repositories. http://roar.eprints.org/, [accessed 19 Aug, 2011].
6) Analysis of comments and implementation of the NIH public access policy. The National Institutes of Health, 2008. http://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf, [accessed 19 Aug, 2011].
7) K.Baba, N.Hoshiko, E.Kudo, N.Yoshimatsu, and E.Ito. Semi-automated paper-registration system for institutional repository. In *The Third International Conference on Awareness Science and Technology*, 2011.
8) K.Baba, E.Ito, and S.Hirokawa. Co-occurrence analysis of access log of institutional repository. In *Japan-Cambodia Joint Symposium on Information Systems and Communication Technology*, pages 25–29, 2011.
9) K.Baba, M.Mori, and E.Ito. A synergistic system of institutional repository and researcher database. In *The Second International Conferences on Advanced Service Computing*, pages 184–188. IARIA, 2010.
10) K.Baba, M.Mori, and E.Ito. Identification of scholarly papers and authors. In *Networked Digital Technologies*, volume 136 of *Communications in Computer and Information Science*, pages 195–202. Springer-Verlag, 2011.
11) K.Baba, M.Mori, E.Ito, and S.Hirokawa. A feedback system on institutional repository. In *The Third International Conference on Resource Intensive Applications and Services*, pages 37–42. IARIA, 2011.
12) K.Baba, T.Tanaka, E.Ishita, M.Mori, E.Ito, and S.Hirokawa. Evaluation of link system between repository and researcher database. In *International Conference on Asia-Pacific Digital Libraries*. Springer-Verlag, 2011.
13) S.Harnad, T.Brody, F.Vallieres, L.Carr, S.Hitchcock, Y.Gingras, C.Oppenheim, H.Stamerjohanns, and E.Hilf. The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4):310–314, 2004.
14) P. Suber. Open access overview. Open Access News, 2007. http://www.earlham.edu/~peters/fos/overview.htm, [accessed 19 Aug, 2011].