ENHANCING THE ELECTRONIC LIBRARY SYSTEM THROUGH THE INTEGRATION OF HETEROGENEOUS DATA SOURCES

HIRONOBU GOTODA, KEIZO OYAMA, and JUN ADACHI

National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Email: {gotoda, oyama, adachi}@nii.ac.jp

ABSTRACT

NACSIS-ELS is an electronic library service of Japanese academic journals provided by National Institute of Informatics (NII). In this paper, we will describe how the functionality of NACSIS-ELS has been enhanced since its initial startup in 1997. NII has been compiling and providing other databases such as CiNii, REO, and Webcat Plus afterward, and the metadata of all these sources are being integrated into a unified system. We will illustrate the details of this system design, and explain the technologies supporting the integration. The issues for further development are also mentioned, in particular, the cooperation with other services.

INTRODUCTION

National Institute of Informatics (NII) is an inter-university research institute in Japan founded in April 2000. NII takes over the development and operation of the infrastructure for disseminating the scholarly information, which National Center for Science Information Systems (NACSIS) had been concerned with for a long time in cooperation with the universities and national institutions in Japan. The Electronic Library Service, called NACSIS-ELS, is an outcome of such activities, from which researchers can retrieve via the Internet the photographic reproductions of academic journals and magazines as well as their bibliographic information [1-3].

As of June 2005, more than 2 million articles are stored in NACSIS-ELS collected by permission from the journals and magazines published by 265 academic societies in Japan. They cover a large proportion of academic papers written in Japanese. The average number of access to NACSIS-ELS is now over 100 thousand for every month. It can be said that NACSIS-ELS has been grown up to be one of the core tools for Japanese researchers when looking for scholarly information. However, there are several shortcomings with NACSIS-ELS. First and foremost, it does not contain any articles written in English. A recent survey has shown that the majority of papers written by Japanese researchers are published in foreign journals [4]. The lack of English contents certainly limits the role of NACSIS-ELS in scholarly communication. Secondly, the type of information provided by NACSIS-ELS is confined to papers. Although journal articles are the primary source of information for most researchers, they frequently refer to books, reports and Web pages as well. Such kind of information is entirely missing in NACSIS-ELS.

To address these problems, we have redesigned the architecture of NACSIS-ELS in coordination with the other information services provided by NII. The newly born NII-ELS is an enhancement of NACSIS-ELS, which works in concert with the NII's Citation Information Service called CiNii [5-6]. Users accessing to CiNii will subsequently be guided to the archives of papers such as NII-ELS and NII-REO (NII Repository of Electronic Journals and Online Publications) [7]. Guiding the users to other information services, such as the digital libraries at foreign publishers' sites, will also be possible if the metadata of these services are supplied to CiNii. The metadata collected from various services are merged at CiNii into a unified database, where the identical records are linked to each other, and similar records are grouped together. As a result, citation links from the papers in NII-ELS to the books or papers in other services will become active, and can be followed with a few clicks. This makes it easier to find related information across the boundaries of services.

In this paper, we will first describe how NACSIS-ELS has been transformed to NII-ELS focusing on the integration of the bibliographic information provided by ELS and the citation information provided by CiNii. We will then discuss the methods and framework for unifying the metadata from multiple data sources. Finally, we will outline the on-going projects aiming at linking between different types of information such as papers, books, and project reports.

OVERVIEW OF NACSIS-ELS

The Electronic Library Service (NACSIS-ELS) is an information service that enables users to retrieve materials from the image databases containing photographic reproductions of academic journals and magazines as well as from the bibliographic databases containing the secondary information of articles and journals [1-3]. Users can search for journal articles by title, author, or keywords from the computers connected to the Internet. They can also select articles from the tables of contents or by browsing through pages. The images of the selected articles can be sent to the local printers to make high-quality printouts of desired pages.

Brief History of NACSIS-ELS

The main contents of NACSIS-ELS are academic journals published by Japanese academic societies. The amount of information available has steadily expanded as the number of participating academic societies increases. At the time the trial service of NACSIS-ELS started in February 1995, only 3 academic societies were providing their journals. The full-fledged service began in April 1997 with a collection of about 10 thousand articles, when 37 academic societies were participating in NACSIS-ELS. The number of participating societies exceeded 100 in 1999, and the number of articles exceeded 1 million in 2001 as shown in Fig.1.

There were several choices to make in designing the NACSIS-ELS. One was to choose between image or full-text as the format of contents. When the conceptual design of NACSIS-ELS was being carried out, the use of SGML, which was considered to be a probable candidate for the standard of full-text documents, was not so popular in Japan. We decided to employ images to represent the main body of articles, which enabled the users to view the contents in the layout of printed documents. On the other hand, the bibliographic information, such as title, author, keywords, and abstract of each article, is stored in full-text in the bibliographic databases. The combination of image and full-text data is one of the remarkable features of NACSIS-ELS.

Although the main contents of NACSIS-ELS are academic journals, it is theoretically possible to hold any type of information accompanied by an appropriate set of metadata. Examples of such information are books collected at conventional libraries. The major obstacles in extending range of information are the issues related to copyright protection. In the initial design of NACSIS-ELS, page images could be viewed only in special browsers or plugins exclusively developed for that purpose. Downloading the images to local files was not permitted. That restriction was loosened a bit as Adobe's PDF became a common format for document delivery and was adopted by NACSIS-ELS.

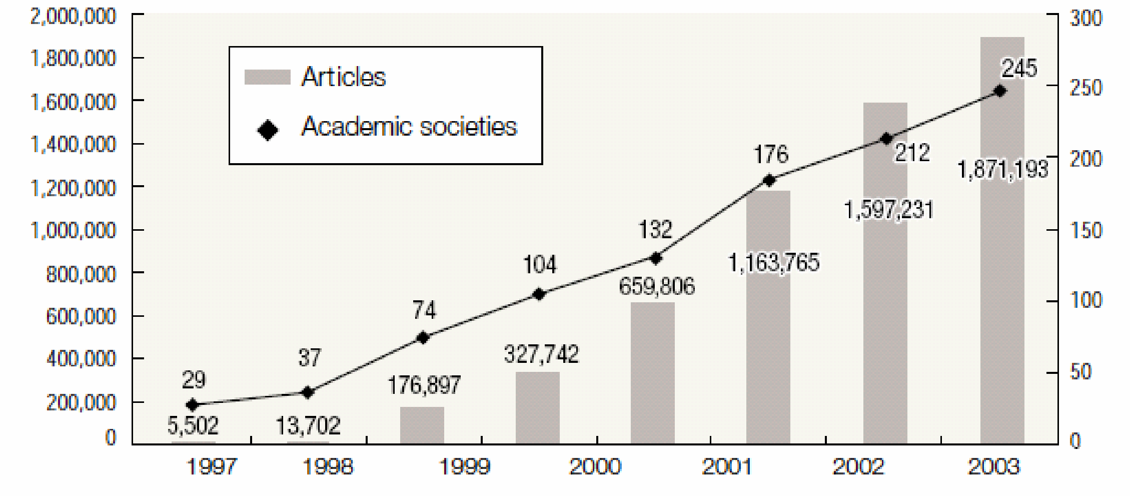


Fig.1 Growth of NACSIS-ELS in terms of the number of articles and societies (As of March 2004).

Problems of NACSIS-ELS

As explained before, we have observed steady growth of contents held by NACSIS-ELS. As of June 2005, more than 200 academic societies are participating in NACSIS-ELS. The number of articles available exceeds 2 million. It is apparent that NACSIS-ELS now plays a substantial role in the circulation and dissemination of scholarly information in Japan.

However, the contents covered by NACSIS-ELS are limited to the articles written in Japanese. There are no foreign publishers or academic societies participating in NACSIS-ELS. It is not likely that the situation will change in the near future, since major foreign publishers such as Elsevier or Springer have their own digital libraries. The lack of English contents can be a fatal drawback of NACSIS-ELS. According to a recent survey on the publication of scholarly papers, almost 80% of scholarly papers written by Japanese researchers are published in foreign journals[4]. This implies that Japanese researchers are paying more attention to foreign journals, which NACSIS-ELS does not cover.

Furthermore, for many researchers, journal articles are not the only source of information. They will also look at books and Web pages to get necessary information. Again, such kind of information is not included in NACSIS-ELS.

The incomplete coverage of contents may lead to an undesirable side effect. As an advanced feature of NACSIS, hyperlinks are embedded in the references of articles, which enable the users to follow the citation links. The targets of these links should be articles also stored in NACSIS-ELS. As the rate of coverage is lowered, there will be less active links embedded in the articles. This may have a negative influence on the behavior of users willing to collect related information.

ENHANCING NACSIS-ELS

We have already seen that the contents currently held in NACSIS-ELS are limited to journal articles published by Japanese academic societies. Due to this limitation, NACSIS-ELS does not necessarily satisfy the needs of Japanese researchers, who also want to access other types of information such as articles published in foreign journals.

Overcoming the Limitation

One way to address the above problem is to import collections of metadata from other information sources, and made them searchable in NACSIS-ELS. Nowadays, information sources such as digital libraries typically include contents as well as metadata. Metadata are often circulated among information providers or stored in secondary databases. By collecting and incorporating such metadata as much as possible, we expect that NACSIS-ELS will become more convenient to use.

However, merely collecting the metadata will not always improve the situation. The format of metadata may differ from source to source. Therefore, it is necessary to convert the collected metadata to normal forms, which is not easy to accomplish. Furthermore, collections of metadata obtained from various sources may include duplicate records. For better usage of the entire collections, such records should be detected and eliminated. This is a kind of “record linkage problem”, whose solution will be discussed later.

The metadata collected from distributed sources should be converted to normal forms, and checked for duplicate entries. We call these processes “metadata integration.” Metadata integration is a very costly task, sometimes requires review by human beings. As such, it should be done offline and separately from the service modules.

Integrating Metadata

NACSIS-ELS is not the sole service that NII is providing for academic communities in Japan. There are several collections whose metadata could be integrated with NACSIS-ELS. One such example is CiNii (Citation Information by NII)[5-6], which provides a gateway to the Citation Database for Japanese Papers (called CJP)[8]. Another example is NII-REO (NII Repository of Electronic Journals and Online Publications), which holds an archive of articles of foreign journals and proceedings[7]. Currently, Oxford University Press, Kluwer Academic Publishers[1], and IEEE Computer Society are supplying their articles to NII-REO based on the contracts with the Japan Association of National University Libraries.

While NACSIS-ELS and NII-REO hold both the primary and secondary information, CiNii holds only secondary information (metadata). CiNii, in its nature, is a navigator who guides the users to pieces

[1] Kluwer Academic Publishers and Springer-Verlag were merged into Springer in July 2004.

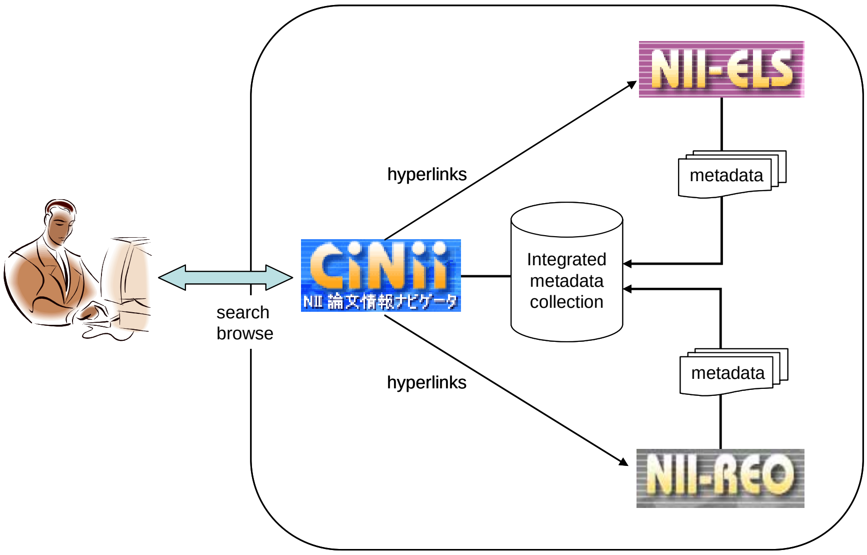


Fig.2. CiNii serving as the front-end of ELS and REO.

of information they are looking for. Having examined the situation, we decided to integrate all metadata into CiNii.

Now CiNii becomes the front-end of ELS or REO (Fig.2). Users can search information at CiNii and follow the links from CiNii to ELS or CiNii to REO. In this framework, functions for searching and browsing, which were included in the original design of NACSIS-ELS, are no longer necessary. So a simplified version of NACSIS-ELS has been designed, which is called NII-ELS. NII-ELS is dedicated to archiving articles written by researchers in Japan. We are now transferring the contents of NACSIS-ELS to NII-ELS. It will be finished by the end of March 2006, when NACSIS-ELS will cease functioning.

Detecting and Eliminating Duplicate Records

The term “record linkage” refers to the art of detecting and eliminating duplicate entries. Record linkage becomes a crucial problem when integrating multiple databases that have been constructed independently. It is also considered to be one of the key issues in integrating heterogeneous Web resources.

Record linkage typically consists of three stages. In the first stage, pairs of records that possibly refer to the same entities are enumerated. The pairs thus enumerated are examined in detail in the second stage. Here each pair is labeled as “match”, “possible match”, or “non-match” based on the matching score calculated using a predefined matching function. Finally, pairs that are labeled as “possible match” are examined by human reviewers. Details of these stages are described in [9].

Table 1. Two bibliographic records that refer to the same article.

1. TAGUCHI Isao *Observation of Nonlinear Conduction in (NbSe_4)_3I* : II. LOW TEMPERATURE PROPERTIES OF SOLIDS : Change Density Waves Japanese Journal of applied physics. Supplement Vol. 26(3.pt1), pp.619-620 (1987)
2. TAGUCHI I Observation of nonlinear conduction in (NbSe4)3I. Jpn J Appl Phys Suppl No.26-3 Pt.1, pp.619-620 (1987)

In our experience, in integrating 1.3million bibliographic records of ELS with 0.6 million records of CJP, we found that about 100 thousand records were redundant. Among the 100 thousand, the pairs of records that were put to human review were about 5 thousand. Although these figures are largely dependent on the nature of databases on target, it can be said that we have succeeded to construct an efficient framework for integrating bibliographic databases.

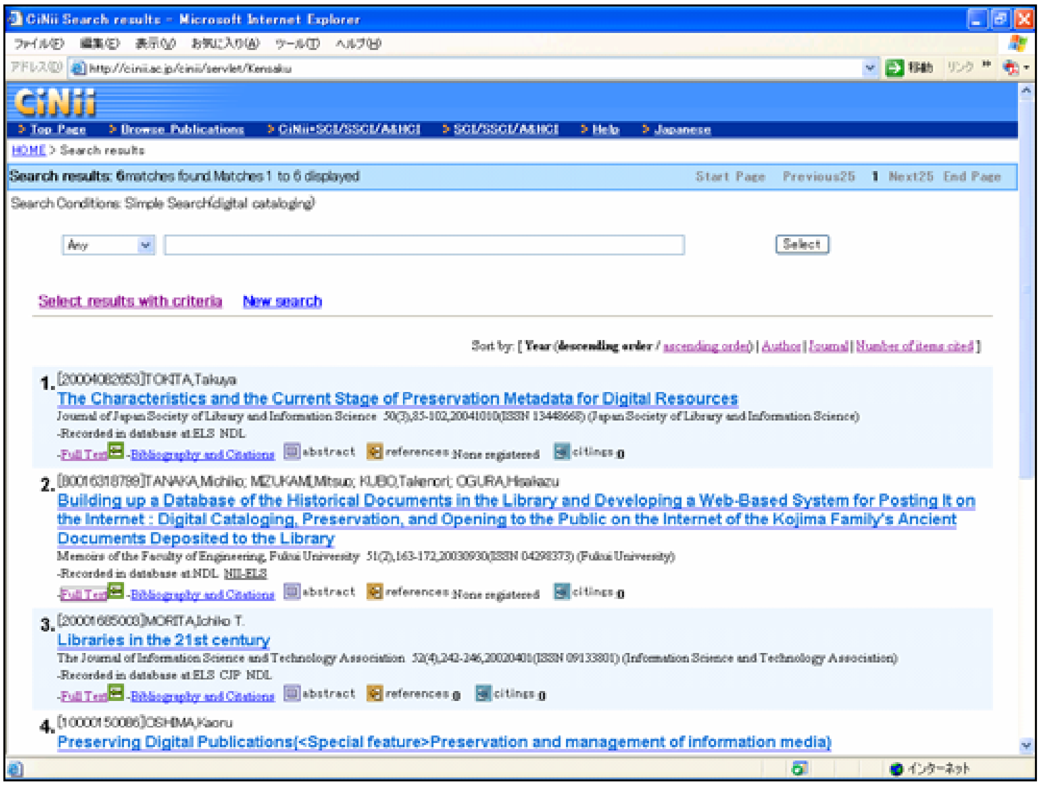
INTEGRATING HETEROGENEOUS SOURCES

The collections of metadata integrated into CiNii are not limited to those of NII-ELS and NII-REO. As of June 2005, Japanese Periodicals Index compiled by the National Diet Library has also been integrated. This is one of the largest index databases extracted from the journals published in Japan, focused on academic journals, research bulletins, and specialist magazines. More are being planned for the future, including Union Catalog Databases (serviced via Webcat Plus[10]) and Project Report Databases (serviced via KAKEN).

Incorporating Union Catalog Databases

The Union Catalog Databases (CAT) record the holdings of books and serials in university libraries and similar institutions. The databases have been created using the cataloging system called NACSIS-CAT, and shared by the librarians. As of March 2005, the number of libraries connected to the system is 1036, with about 80 million records. The databases can be accessed on the World Wide Web through the Webcat Plus service.

We are planning to integrate CAT with the Citation Database for Japanese Papers (CJP). If this integration is completed, hyperlinks are established between references of articles records of holdings in the catalog databases. Users can follow such links and find out which libraries hold the books referred to in a particular article.


CiNii Search results - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://ci.nii.ac.jp/cinii/servlet/Kensaku
移動 リンク
CiNii
Top Page
Browse Publications
CiNii+SCI/SSCI/A&HCI
SCI/SSCI/A&HCI
Help
Japanese
HOME > Search results
Start Page Previous25 1 Next25 End Page
Any
Select
Select results with criteria
New search
インターネット

Search results: 6matches found Matches 1 to 6 displayed

Search Conditions: Simple Search(digital cataloging)

Sort by: [Year (descending order / ascending order) | Author | Journal | Number of items cited]

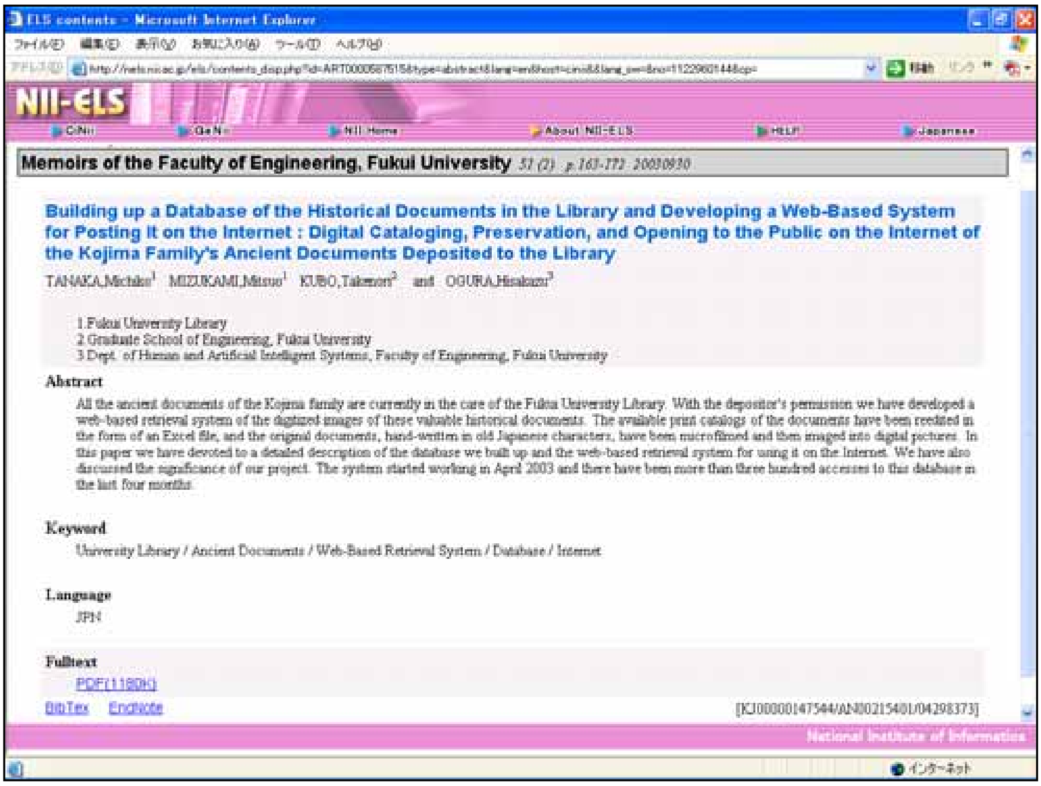
1. [20004082653]TOKITA,Takuya
The Characteristics and the Current Stage of Preservation Metadata for Digital Resources
Journal of Japan Society of Library and Information Science 50(3),85-102,20041010(ISSN 13448668) (Japan Society of Library and Information Science)
-Recorded in database at ELS NDL
-Full Text -Bibliography and Citations abstract references None registered citings 0

2. [80016318799]TANAKA,Michiko; MIZUKAMI,Mitsuo; KUBO,Takenori; OGURA,Hisakazu
Building up a Database of the Historical Documents in the Library and Developing a Web-Based System for Posting It on the Internet : Digital Cataloging, Preservation, and Opening to the Public on the Internet of the Kojima Family's Ancient Documents Deposited to the Library
Memoirs of the Faculty of Engineering, Fukui University 51(2),163-172,20030930(ISSN 04298373) (Fukui University)
-Recorded in database at NDL NII-ELS
-Full Text -Bibliography and Citations abstract references None registered citings 0

3. [20001685003]MORITA,Ichiko T.
Libraries in the 21st century
The Journal of Information Science and Technology Association 52(4),242-246,20020401(ISSN 09133801) (Information Science and Technology Association)
-Recorded in database at ELS CJP NDL
-Full Text -Bibliography and Citations abstract references 0 citings 0

4. [10000150086]OSHIMA,Kaoru
Preserving Digital Publications(<Special feature>Preservation and management of information media)

Fig. 3. A snapshot of CiNii. By clicking the “Full Text” link of the second article, you will get the page shown in Fig.4.


ELS contents - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://nels.nii.ac.jp/els/contents_disp.php?id=ART0000587515&type=abstract&lang=en&host=cinii&lang_sw=&no=1122960144&cp=
移動
リンク
NII-ELS
CiNii
GeNii
NII Home
About NII-ELS
HELP
Japanese
National Institute of Informatics
インターネット

Memoirs of the Faculty of Engineering, Fukui University *51 (2) p.163-172 20030930*

Building up a Database of the Historical Documents in the Library and Developing a Web-Based System for Posting it on the Internet : Digital Cataloging, Preservation, and Opening to the Public on the Internet of the Kojima Family's Ancient Documents Deposited to the Library

TANAKA,Michiko[1] MIZUKAMI,Mitsuo[1] KUBO,Takenori[2] and OGURA,Hisakazu[3]

1.Fukui University Library
2.Graduate School of Engineering, Fukui University
3.Dept. of Human and Artificial Intelligent Systems, Faculty of Engineering, Fukui University

Abstract

All the ancient documents of the Kojima family are currently in the care of the Fukui University Library. With the depositor's permission we have developed a web-based retrieval system of the digitized images of these valuable historical documents. The available print catalogs of the documents have been reedited in the form of an Excel file, and the original documents, hand-written in old Japanese characters, have been microfilmed and then imaged into digital pictures. In this paper we have devoted to a detailed description of the database we built up and the web-based retrieval system for using it on the Internet. We have also discussed the significance of our project. The system started working in April 2003 and there have been more than three hundred accesses to this database in the last four months.

Keyword

University Library / Ancient Documents / Web-Based Retrieval System / Database / Internet

Language

JPN

Fulltext

PDF(1180K)

BibTex EndNote [KJ00000147544/AN00215401/04298373]

Fig. 4. A snapshot of NII-ELS.

The integration will be done based on fields of author, title, publisher, and year of publication. The references in CJP may be in abbreviated while complete and well-formatted information is stored in CAT. Finding identical records between CJP and CAT would be somewhat more difficult than that of CJP and ELS or REO.

Incorporating Project Report Databases

The Project Report Databases serviced through KAKEN are the compiled records of projects supported by Grant-in-Aid for Scientific Research sponsored by the Ministry of Education, Culture, Sports, Science and Technology (MEXT). Grants are awarded to projects organized by individual researcher or research groups at Japanese universities or research institutes. Each report consists of the title of the project, names and affiliations of researchers, a summary of achievements, and a list of publications. As of June 2005, about 300 thousand reports are held in the databases.

We view that KAKEN is useful for identifying researchers or research groups. Using such information, articles published by the same authors can be grouped together, which will help the users of CiNii to find related information.

The key to this attempt is a reliable method for matching the publication lists in KAKEN databases to the integrated collections of metadata of CiNii. Since the publication lists in KAKEN databases are not well-formatted, we anticipate that the quality of computer-automated matching is not so high.

CONCLUSIONS

National Institute of Informatics has been providing various services to academic communities in Japan. These include electronic library service (ELS), union catalog service (Webcat Plus), repository of electronic journals and online publications (REO), and citation information service (CiNii). Recently, redesigning or modification of these services has been carried out so that they will work in concert. In particular, NACSIS-ELS has been simplified NII-ELS, and collections of metadata in ELS has been integrated into CiNii. We have described the background and outcome of this simplification and integration.

We will continue to integrate more and more collections of metadata into CiNii. Some of the on-going projects have also been outlined aiming at integrating CiNii and CAT or KAKEN.

REFERENCES

[1] Adachi, J. “NACSIS electronic library system: Its design and implementation,” *Proc. International Symposium on Digital Libraries (ISDL95)*, pp.36-41, 1995.

[2] Adachi, J. “Digital library system of document images focusing on metadata (in Japanese),” *Transactions of IEICE*, J84-D-I (6), pp.768-776, 2001.

[3] NACSIS-ELS homepage [http://www.nii.ac.jp/els/els-e.html] (last access 2005.7.1)

[4] Adachi, J., Negishi, M., Tsuchiya, S., Konishi, K., Oba, K. and Okumura, S. “Publishing role in the scholarly communication. Dispatch of Japanese scientific research seen in SPARC/JAPAN and University Library (in Japanese)”, *The Journal of Information Science and Technology*, 53(9), pp.429-434, 2003.

[5] CiNii Home [http://ci.nii.ac.jp/] (last access 2005.7.1)

[6] Oyama, K., Aizawa, A., Gotoda, H., Kojin, S. and Otsuna, K. “Development of scholarly article information navigator (in Japanese)”, *IPSJ SIG Technical Report*, FI-75, pp.119-126, 2004.

[7] NII-REO [http://reo.nii.ac.jp/] (last access 2005.7.1)

[8] Negishi, M., Sun, Y. and Shigi, K. “Citation database for Japanese papers: A new bibliometric tool for Japanese academic society,” *Scientometrics*, 60(3), pp.333-351, 2004.

[9] Aizawa, A. and Oyama, K. “A fast linkage detection scheme for multi-source information integration,” *Proc. International Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2005)*, pp.31-40, 2005.

[10] Webcat Plus [http://webcatplus.nii.ac.jp/] (last access 2005.7.1)