京 都 大 学 防 災 研 究 所 年 報　第47号 C 平成16年3月
Annuals of Disas. Prev. Res. Inst., Kyoto Univ., No. 47 C, 2004

Digitization of disaster prevention Printed Matters and Video information,and construction of the search engine which can search these on the Internet website at high speed

Hideki MATSUURA, Kenichi TATSUMI, Hideo TAGAWA,
Yoshinori YOSHIDA, Tsutomu MIURA, Tetsuro TAKAYAMA,
Hiroo WADA and Norio HIRANO

Synopsis

At Disaster Prevention Research Institute (DPRI), a large amount of disaster prevention research base Printed Matters has stored since 1951. The aim of assignment research subject of Division of Technical Affairs is that these base materials are widely exhibited to society, and can be used. In this paper, we show digitization of the DPRI Annuals that is published every year, and construction of the search engine which can search the Printed Matters on the Internet website at high speed in fiscal year 2003. Much workforce and time was needed for the digitization work. We reduced a document and picture size of the Annuals to the minimum, which were comprehensible. In addition, we devised so that the work efficiency was enhanced. As a result, about 20,000 pages of the DPRI Annuals (12 years) were exhibited to the public on Internet website in a short period. The three search engine services, which users can be selected, has been available on the Internet website in May 2003. Anyone enabled the above ingenuity to browse the Annual documents easily, and the search engine that can respond to a large user layer was realized.

Keywords: DPRI Annuals; OCR; Digital image; PDF; search engine; database

1. Introduction

Disaster Prevention Research Institute (DPRI) was established in 1951. The field of research in DPRI now covers a wide range of disaster-related topics. DPRI plays a key role in the research on "Investigation of disaster theory and construction related to disaster prevention study".

Recently, telecommunication and multimedia technologies have developed rapidly. As a result, much information is exchanged all over the world. Therefore, to retrieve only the most necessary information and effectively use the information is important.

To date, DPRI has accumulated a large amount of disaster prevention materials. To further promote disaster prevention research, dissemination of Printed Matters in these materials is necessary. If these Printed Matters were available on the DPRI Internet website, that would be informative to the public.

We paid particular attention to the DPRI Annuals that are an important component of Printed

Matters. After the first publication in 1958, the Annuals continued to be published up to now. However, obtaining the old Annuals is difficult because the reprint of the old Annuals is difficult. The data of disaster prevention in the Annuals is not readily available, when the public require the data.

Therefore, we wrote this paper to the methods to exhibit the Annual documents on Internet website, and to enable the users to acquire the documents at any time. Specifically, the Annual paper documents were converted to the Portable Document Format (PDF) files, and were exhibited on the DPRI Internet website. With that three search engine were constructed, which is usable for the public to acquire necessary data in the DPRI Annuals.

Note that a particular bit of software (Acrobat Reader) is required to open the PDF file. Acrobat Reader can be downloaded for free on the website of Adobe Systems Incorporated Company. Because PDF format is a *de facto* standard for electronic documents in Japan, we chose the PDF format.

2. Digitizing the DPRI Annual documents

We started to digitize the Annuals from No.45.

2.1 Processes of digitizing the Annual papers

The sequence of digitizing the Annual papers is outlined below:

[1] Each page of the papers is converted into digital image (bit map image format) files on an image-scanner and is saved.

[2] The blots of digital image files are eliminated on the Picture edit software.

[3] Digital image files made in process [1] and [2] are converted to PDF files on WinReader8.0 (a type of Windows OCR application software).

[4] The character strings of the Annual documents are extracted from digital image files on WinReader8.0. The strings are saved in the text file. In addition, we input strings data into a database. Process [4] was done simultaneously with process [3].

Most work was relatively simple, however much workforce was needed. Therefore, we trained part-time staffs, who supported us the work from January to March, 2003. If the work of these processes is done in order as they are, it is should be emphasized that great time will be spent on the work.

Hence the processes were divided into as smallest possible tasks, for the increase in efficiency of time shortening and the work sake. Consequently, there were not great differences between individuals in the outcome of the work. The operational flowchart is shown in Figure 1.

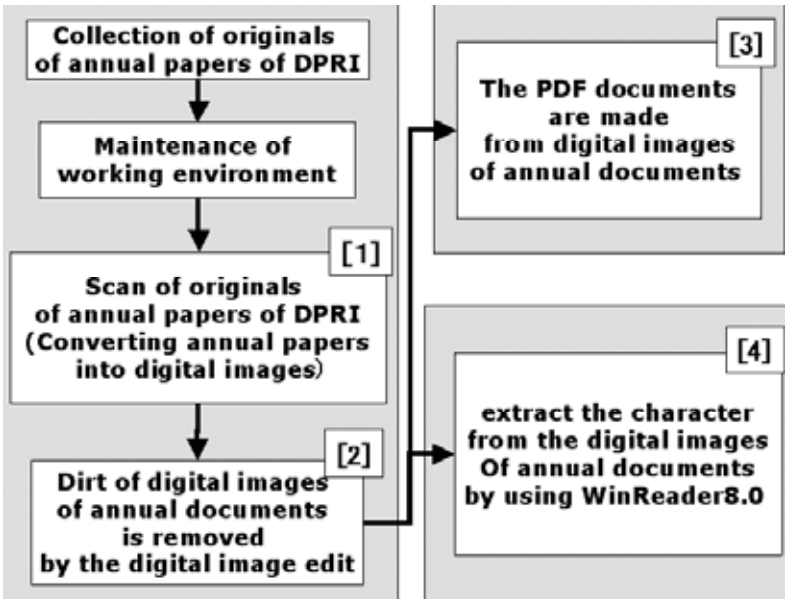

Fig. 1 Digitizing DPRI Annual documents

In addition, technical officials did the more complex part of the processes that was required professional skills. The three specific works are as follows:

- Supervision of part-time staffs' work and management of the saved data
- Standardization of the format of digital image data in the process [2]
- Making the website for exhibiting the Annuals and constructing the DPRI Annuals search engine

2.2 PDF documents of the DPRI Annuals

In the process [3], the character strings of the Annual documents with digital images were able to be converted to PDF file. However, for "Loading PDF file error", there may be one problem that PDF files, which contain Japanese fonts, could not be browsed in foreign countries. The following two methods were employed to solve this problem:

[A] The Japanese fonts are built into a PDF file.

[B] Only digital images are converted to PDF file.

Method [A] requires time and cost very much, hence we chose method [B] to make PDF files composed only of digital images. The work described herein was conducted from January to March 2003. As a result, about 20,000 pages of the DPRI Annuals (12 years) are exhibited to the public on Internet website. The search engine service that offers data of the DPRI Annuals on the Internet website began in May 2003.

3. Construction of the search engines for browsing the DPRI Annuals

The search engine for browsing the DPRI Annuals that we constructed this time consists of three components:

- The Index search: Users choose the volume number of the DPRI Annuals and search the Annual document from the displayed indexes.
- The Category search: Users search the Annual document using the information of the limited categories which users know (title, author, publication year and other information).
- The Full-text search: Users search the Annual document using arbitrary words.

We constructed the three search engines to search data of the Annuals to be efficient and convenient for users. Users' knowledge related to DPRI Annuals varies depending on their backgrounds. Thus, we enabled the users to select the most appropriate search engine service for users.

3.1 The index search engine

The index search engine is directed to the user who has accurate knowledge about the Annuals such as publication year and other information. The list of volumes (Fig. 2) and indexes of each volume (Fig. 3) were written in HTML files. In the HTML file of indexes, the link to a PDF file of each Annual document (Fig. 4) is specified by the document title. After the document is found in indexes, users can click the title that users wish to access and browse a PDF file of the Annuals.

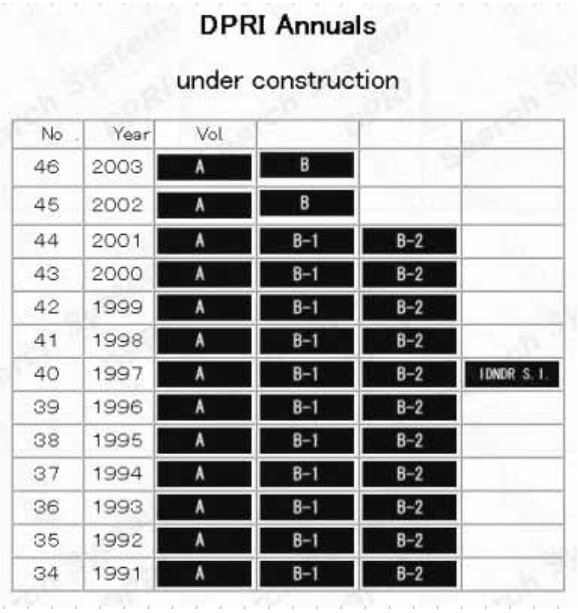

Fig. 2 DPRI Annuals list

DPRI Annuals No.43 (2000) Index

Volume B-1

Fig. 3 DPRI Annuals indexes

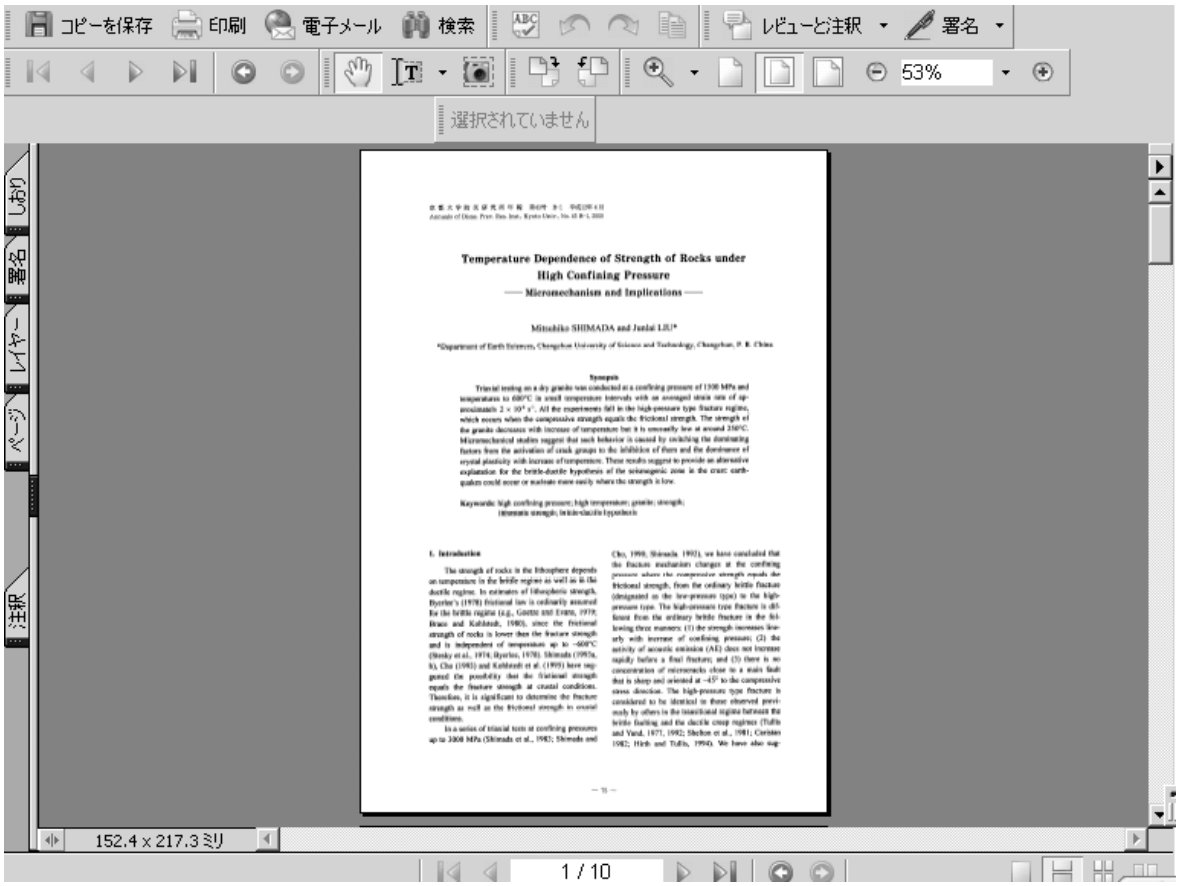

Fig. 4 A PDF file of DPRI Annuals

3.2 The category search engine

If users don't know in which volume the Annual document that users wish to browse, users may be unable to search the document with the index search.

We constructed the category search engine so that users can search the Annual document with information of categories such as titles, authors, synopsis, etc.

The system structure is "Apache+PHP +MySQL", the soft are the open source software.

MySQL is easy to use compared with expensive "business" databases like Oracle or SQLServer. The server script language "PHP" which can directly connected to the database operates as a module of Apache. The structure has high affinity with database, and has smaller burdens than the CGI, which runs with a different process. The system structure of database "Apache+PHP+MySQL" which is combined both of them, is supported throughout the world.

We chose "Apache+PHP+MySQL" structure because this structure has an excellent cost performance and has special features (high speed and stability). Note that PDF files of the DPRI Annuals contain only digital images. Thus, we extracted the strings data of Annual document related to categories and compiled the data into Microsoft Access database. The Microsoft Access database is specified interface parameters with MySQL ODBC Driver to be exported to the MySQL database.

To use the category search engine, users have to write arbitrary words in each text box that users can type the category keyword into, and push the "search" button on the search form of category search website (Fig. 5). Then users select the Annual document that users wish to browse among the search results and push "pdfview" in title part (Fig. 6). Consequently, a pdf file of the Annuals PDF document appears.

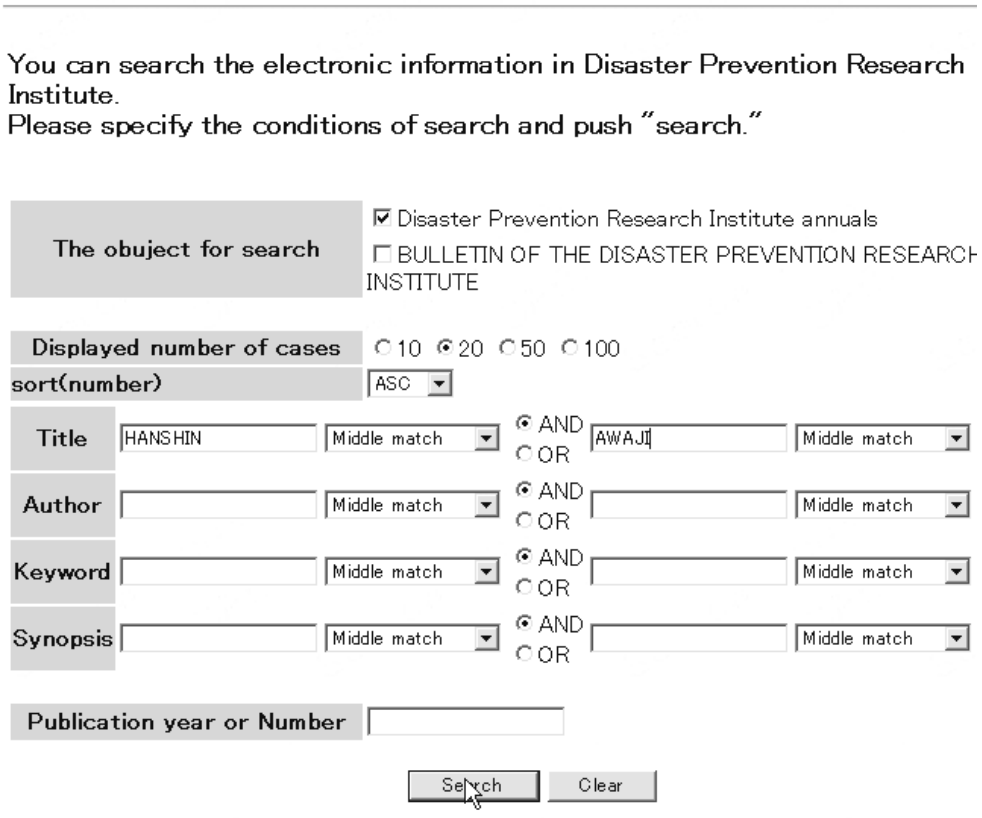

Fig. 5 The category Search engine

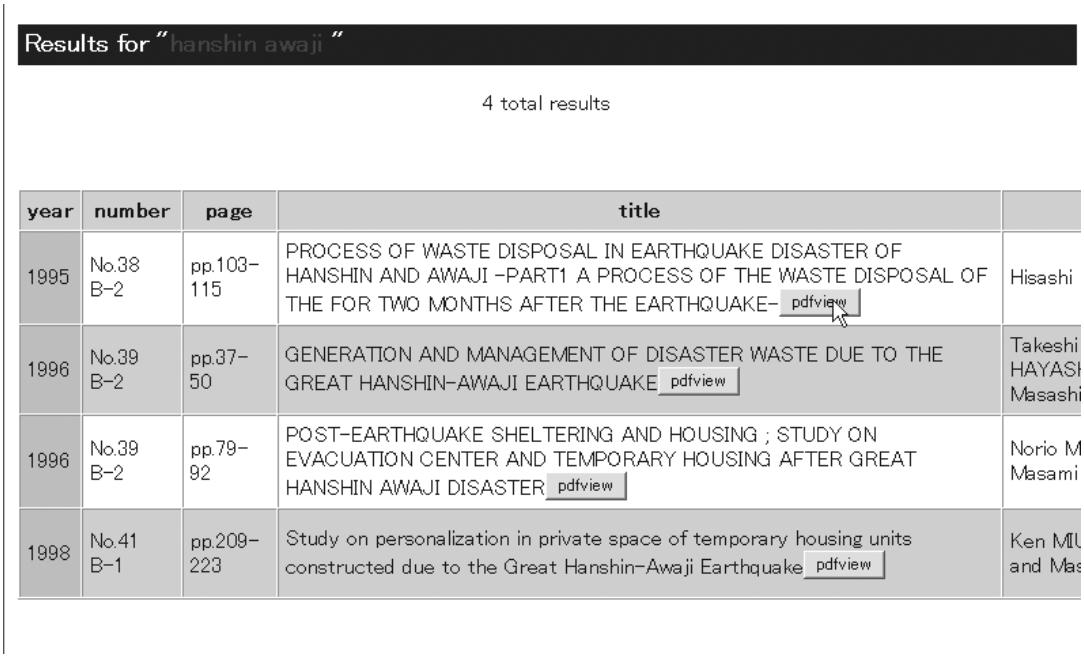

Fig. 6 Result of a search

The workflow of the category search engine is shown in Fig.7. A part of "query" receives the keyword parameter that users typed from the search form of category search engine website, and searches the keyword in the database table "Annual". The database table employs "ID" (Table 1).

To extract only the data that fits specific conditions, we use "SELECT" command; besides, for more obscure conditions search, we use "LIKE" command. In such search, fields of titles, authors, key words, the synopsis, publication years and volume numbers are searched. In addition, records that include some keywords carried from the search form are extracted.

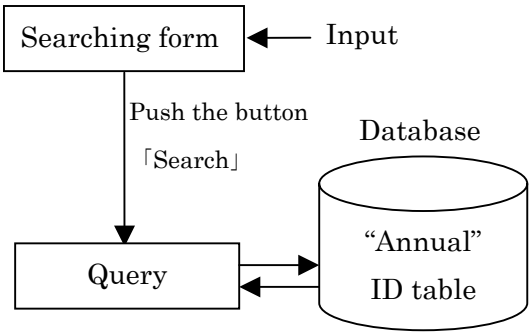

Fig. 7 The workflow of category search engine

Table.1 Field of ID table

Field Name	Data type
ID	Auto Number
Year	Int(11)
Titlej	Mediumtext
Authorj	Mediumtext
Keywordj	Mediumtext
Synopsisj	Mediumtext
Title	Mediumtext
Author	Mediumtext
Keyword	Mediumtext
Synopsis	Mediumtext
Numberj	Varchar(50)
Number	Varchar(50)
Page	Varchar(50)
Link	Varchar(50)
Link2	Varchar(50)

3.3 The full-text search engine

If users have little knowledge of the category keywords that related to the Annual document, users take time to find one in the index search engine or in the category search engine.

Thus, we constructed the full-text search engine that enables them to retrieve with a few free words. Structure of the search engine is “Namazu”, the only open source software in the Japanese full-text search softwares. “Namazu” is a free and easy-to-use Japanese full-text search engine. If “Namazu” runs as a CGI, the search engine can be constructed. Note that several softwares such as “Apache”, “Perl”, “File-Mmagic”, “nkf”, and “KAKASI” are required for the full-text search on the Internet website. The soft wares are open source soft wares and have good cost performance.

Note that the PDF files of the Annual documents contain no text data. Accordingly we made Link-HTML files (reference file that refers to a corresponding PDF file of a Annual document) that added full-text of the Annuals document. Accessing the Link-HTML, users can browse the PDF files of the Annual document after a few seconds. The flow of the search engine is shown in Fig. 8.

To use the full-text search engine, users have to write free-word in the blank space (“Input text box”), and push the “search!” button (Fig. 9). Then users select the Annual document that users wish to browse, and push the document title or "Link is here" in the search result list (Fig. 10). Consequently, a PDF file of the Annuals PDF document appears.

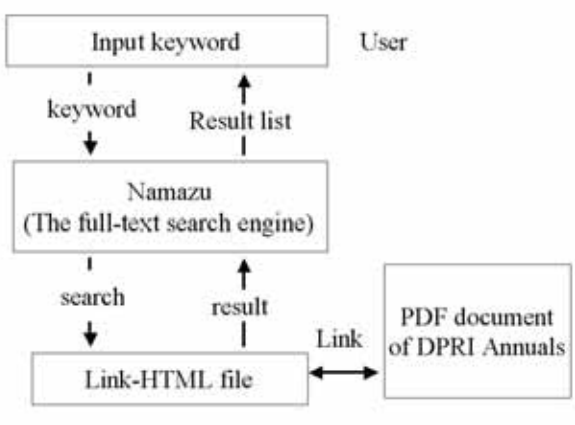

Fig. 8 The full-text search engine

Fig. 9 Input Keywords

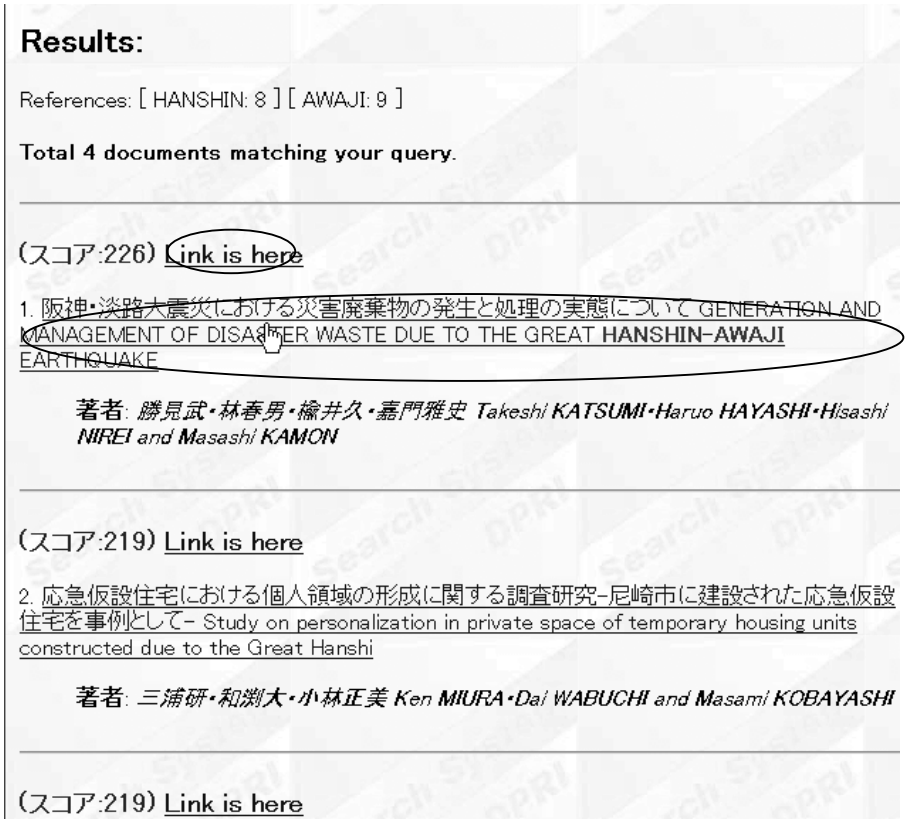

Fig. 10 The full-text search result list

4. Conclusions

By the end of fiscal year 2003, we completed digitization the DPRI Annuals from No.34 to No.45. The service of the three search engines that enable high-speed search of the Annuals has run since May 2003.

In fiscal year 2004, we are going to continue the work of digitization of the remaining Annuals documents and construct upgraded search engine that offer the data of many other disaster prevention Printed Matters in DPRI (Video, Printed matter exceeding A4 size).

Acknowledgments

This research was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) 21st Century COE Program for DPRI, Kyoto University (Program Leader: Prof. Yoshiaki KAWATA, Assignment research Leader: Prof. Haruo HAYASHI) as well as the Grant-in-Aid for Scientific Research (Principal Investigator: Norio HIRANO, Kyoto University).

The authors wish to thank the colleagues and professors that helped us on this project and the 21st Century COE Program committee for allowing us to research on this project.

References

Satoshi TATEOKA, (2003): *A guide to a formal WeB Database system by MySQL+PHP -Point of Web application development Advanced server side programming* -, Gijutsu-Hyohron co.

Hajime BABA, (2001): *Construction and practical use of a system - Japanese full-text search thoroughness guide*-, SOFTBANK co.

Haruo HAYAMI, (2002): *Database IT Text*, Ohmsha co.

要 旨

防災研究所が設立されて以来、研究成果として膨大な印刷物が蓄積されてきた。本稿では、印刷物の中でも毎年発刊される年報の電子ファイル化と、インターネット上で年報文献を高速検索できるシステム構築について、平成15年度分を報告する。文献の電子ファイル化作業は、多大な労力と時間を必要とするが、作業効率をあげることで短期間に2万ページの文献を電子ファイル化した。また、ユーザが用途に応じて検索方法を選択できる、より広い利用者層に対応した年報検索サービスを実現させた。

キーワード: 防災研究所年報; 光学式読取装置; デジタル画像; PDF;検索システム; データベース