# 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 (2)

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

## Synopsis

A large amount of disaster prevention research base Printed Matters has stored since 1951 at the Disaster Prevention Research Institute (DPRI). 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. However, the materials are not the one accumulated on the assumption of being able to use it on Web in the future. Thereby, it is difficult for us to pick up only necessary information of the materials. Accordingly, it is necessary to offer the comprehensible materials on the Internet to advance the disaster prevention study research.

In this paper, we provide method of converting the materials of the DPRI into digitalized materials and construction of the high-speed system that can search the information of disaster prevention on the Web. In last year, we converted the 34-45 Annual papers to the Portable Document Format (PDF) files, which were exhibited on the DPRI Internet website. Moreover, we constructed the three kinds of search engines to search data of the DPRI Annual documents for users efficiently and conveniently. For this year, we show the faster-than-ever-before digitization of the disaster prevention research base materials of the DPRI including the DPRI Annuals. Additionally we show the construction of the search engine which can quickly search the materials on the Internet website.

**Keywords:** DPRI Annuals; OCR; Digital image; PDF; a-high-speed scanner; search engine;

## 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 technology and multimedia technology 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. In particular, we expect that the disaster prevention sources will be more useful for the society, if a lot of the disaster prevention materials of DPRI were exhibited comprehensible on the Internet homepage.

After the first issue of annuals was published in 1958, they keep being published until today.

However, most of the preserved literature in the DPRI is a paper medium.

In the last two years, we completely digitized the DPRI Annuals from No.25 to No.45. However, we find out that "Time" and "Labor" costs of the system for digitalization, which we showed last year, were very high.

Therefore, it is difficult to digitalize the DPRI Annuals of the remainder of 24 years by using the work system in one year. In addition, it is difficult for us to search the disaster prevention information except for a literature. Utilizing a high-speed scanner and the batch processing, we solved “Time” and “Labor” costs problems. And we show the suggestion of the new search engine that search a variety of information for the disaster prevention.

## 2. Digitizing the paper literature

### 2.1 Procedures digitized the Printed literature

The procedure that is digitizing the paper literature can divide into three parts.

The procedures of digitizing the paper literature in the last year:

[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 with the digital image file editor by the manual work.

[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).

Note that the procedure was necessary for a great human effort, if we would work by the past way. Accordingly, we have improved the work system with a high-speed scanner for the increase in the work efficiency.

### 2.2 Improvements of the work efficiency digitized the Printed literature smoothly

A high-speed scanner is a scanner that can convert the paper material into digital image (JPEG image format) files at high speed. For the last a few years, a high-speed scanner is miniaturized, and is becoming more powerful and cheaper. In this research, we used the "Scan Snap" scanner. The "Scan Snap" scanner is more powerful and cheaper than all the other standard scanners on store-bought scanners. The procedures of making into the digitized the literature with this high-speed scanner are as follows.

The procedures of digitizing the paper literature in this year:

[1] Each page of the Annual-papers is converted into digital image (JPEG image format) files by a high-speed scanner “Scan snap” and is saved.

[2] The blots of the digital image file are eliminated on digital image file editor automatically except in some of the manual work.

[3] The digital-picture-files (hereinafter called image pages) are made by procedures [1] and [2], are processed to PDF files by each document on Win Reader 8.0 (a type of Windows OCR application software).

The work efficiency of [1] went up by about five times by introduction of the high-speed scanner compared with the past. However, the digital images that were scanned through the “Scan snap” were yellowish cast because of low degree of brightness and contrast. If the PDF files would be made from those images, the PDF files were stained badly.

We corrected the brightness and contrast on the images using the batch image processing software “Irfan” for efficiency improvement, as shown in Figure 1. Note that we could not correct the brightness and contrast by the batch image software “Irfan” on a part of the digital images, which include "Figure and table". Therefore, we had to manually adjust in case of the pages including “Figure or table” by the image-editing software “ Paint shop”. Furthermore, there were a lot of blots around the edge of each scanned image. We had removed these blots by the manual work in the last year. In this year, we had removed each page of them by the batch image software “Trim Version 1.3” that cuts out only necessary part, as shown in Figure 2. In addition, we extract the textual information from the digital image files with OCR software “WinReader8.0”, so that we construct the search engine that can search for the data files of the disaster prevention research on the Internet website. And we preserved the textual information into the database, which was used for the search engine. Note that we contrived easy ways to reduce the skill difference in the operation whoever did them through the process (1),(2) and (3). We utilized a slimmer batch processing and an assembly-line system.

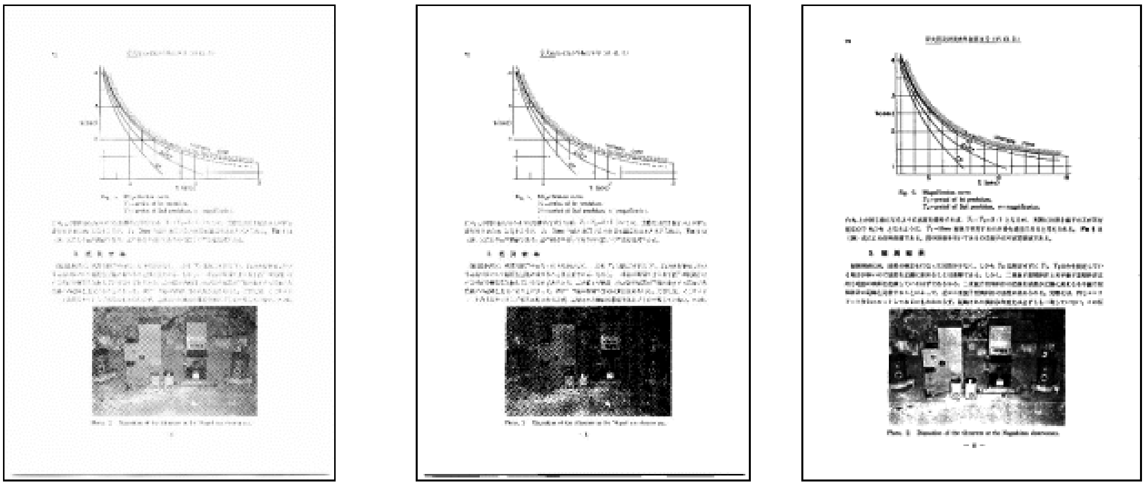


Fig. 1 The image correction of the digital image

(Immediately after Scaning, after an image is automatically corrected, after an image manually corrected, are shown from the left)

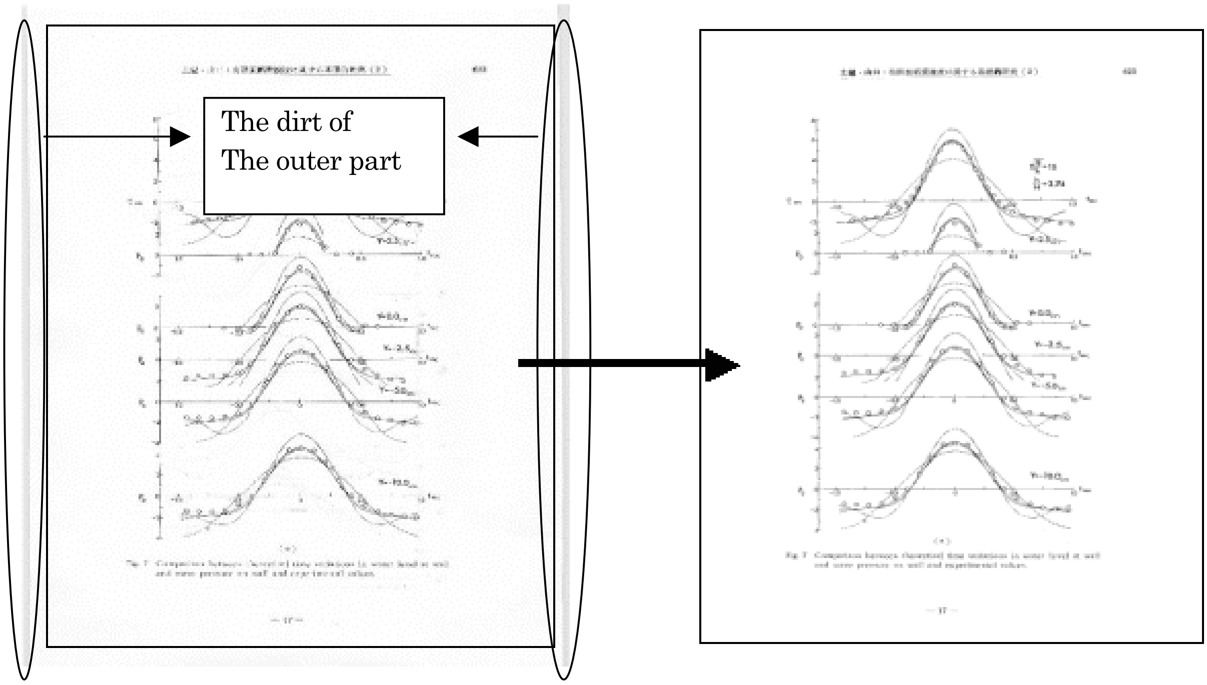


Fig. 2 To remove the dirt of the outer part by the batch digital image trim software "Trim Version 1.3"

As a result, the batch software enabled us to shorten the working hours to correct. Figure 3 shows the specification of a high-speed scanner “Scansnap”.

Finally, rest of the Annual paper documents were converted into the PDF files to exhibit on the DPRI Internet website with the OCR software “WinReader8.0”, the same as last year.

Figure 4 shows the assembly-line work system that can convert the paper published literature into the digitized literature.

| | | |
|---|---|---|
| Description of product | | ScanSnap fi-5110EOX2 |
| Reading system | | Automatic paper feed<br>(Both sides coinstantaneously reading) |
| Reading mode | | Color / Black and white / Auto<br>（Color ・ black and white automatic identification) |
| Optical system / Light source | | Aspheric lens attrition optical system CCD |
| Read rate<br>[It is possible to preserve it with PDF and JPEG file.] | Normal mode | Color 150dpi、Black and white 300dpi:<br>Both sides・Single 15 pages/ minute |
| | Fine mode | Color 200dpi、Black and white 400dpi:<br>Both sides・Single 10 pages/ minute |
| | Super Fine mode | Color 300dpi、Black and white 600dpi:<br>Both sides・Single 5 pages/ minute |
| | Excellent mode | Color 600dpi、Black and white 1,200dpi:<br>Both sides・Single 0.5 pages/ minute |

Fig. 3 the specification of a high-speed scanner “Scansnap”

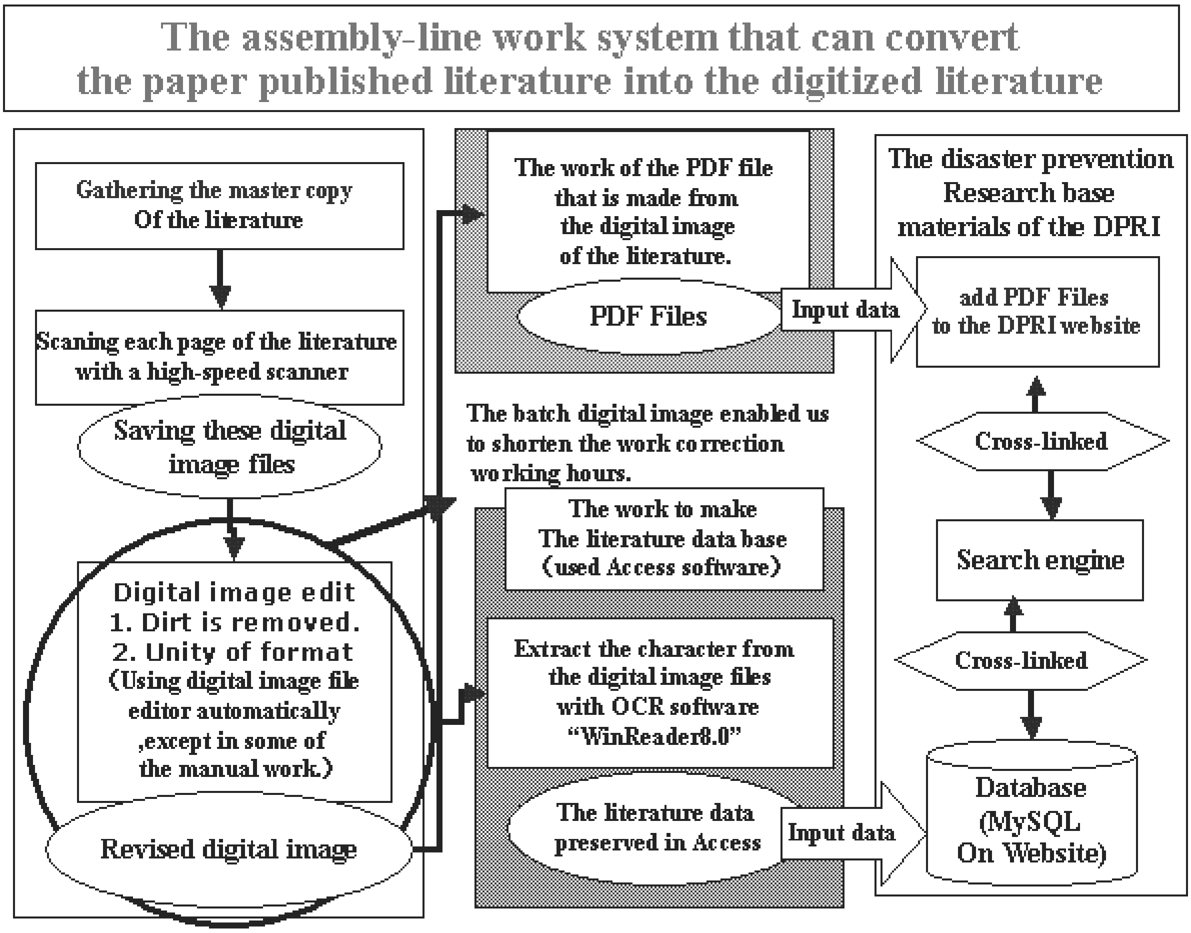

The assembly-line work system that can convert the paper published literature into the digitized literature
Gathering the master copy Of the literature
Scaning each page of the literature with a high-speed scanner
Saving these digital image files
Digital image edit
1. Dirt is removed.
2. Unity of format
（Using digital image file editor automatically ,except in some of the manual work.）
Revised digital image
The work of the PDF file that is made from the digital image of the literature.
PDF Files
Input data
The batch digital image enabled us to shorten the work correction working hours.
The work to make The literature data base （used Access software）
Extract the character from the digital image files with OCR software “WinReader8.0”
The literature data preserved in Access
Input data
The disaster prevention Research base materials of the DPRI
add PDF Files to the DPRI website
Cross-linked
Search engine
Cross-linked
Database (MySQL On Website)

Fig. 4 The assembly-line work system that can convert the paper published literature into the digitized literature

## 3. The disaster prevention research base materials of the DPRI including the DPRI Annuals

### 3.1 Construction of the search engines for browsing the DPRI Annuals in the last year

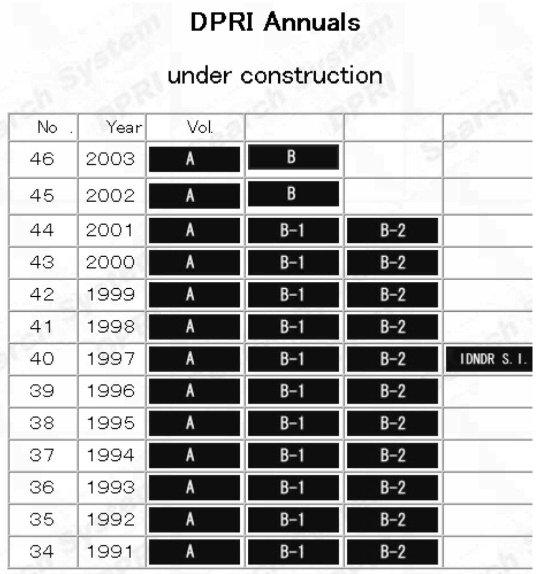
The annual report document search engine that we constructed last year is the following three.

**The Contents Search:** User choose the volume number of the DPRI Annuals and pick up User want a document from the displayed contents.

**The Category Search:** The Category Search: User put the information already user knows in the limited category filed box (title, author, publication year and other information) on the web.

**The Full-text Search:** User put any arbitrary word in the text field box.

The contents search engine in three of the search engines is directed to the user who has accurate knowledge about the Annuals such as publication year and other information. The list of volumes (Fig. 5) and contents of each volume were written in HTML files. On the HTML file of DPRI Annual contents, the link to a PDF file of each Annual document (Fig. 6) is specify with each title. After the document is found in contents, users can access and browse a PDF file that users want to see with a click.

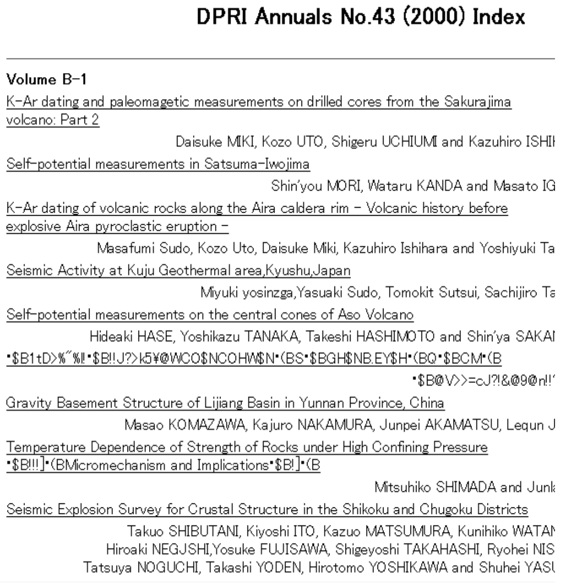


DPRI Annuals

under construction

| No . | Year | Vol. | | | |
|---|---|---|---|---|---|
| 46 | 2003 | A | B | | |
| 45 | 2002 | A | B | | |
| 44 | 2001 | A | B-1 | B-2 | |
| 43 | 2000 | A | B-1 | B-2 | |
| 42 | 1999 | A | B-1 | B-2 | |
| 41 | 1998 | A | B-1 | B-2 | |
| 40 | 1997 | A | B-1 | B-2 | IDNDR S. I. |
| 39 | 1996 | A | B-1 | B-2 | |
| 38 | 1995 | A | B-1 | B-2 | |
| 37 | 1994 | A | B-1 | B-2 | |
| 36 | 1993 | A | B-1 | B-2 | |
| 35 | 1992 | A | B-1 | B-2 | |
| 34 | 1991 | A | B-1 | B-2 | |

Fig. 5 The DPRI Annuals lists



DPRI Annuals No.43 (2000) Index

Volume B-1

K-Ar dating and paleomagetic measurements on drilled cores from the Sakurajima volcano: Part 2

Daisuke MIKI, Kozo UTO, Shigeru UCHIUMI and Kazuhiro ISHI

Self-potential measurements in Satsuma-Iwojima

Shin'you MORI, Wataru KANDA and Masato IG

K-Ar dating of volcanic rocks along the Aira caldera rim - Volcanic history before explosive Aira pyroclastic eruption -

Masafumi Sudo, Kozo Uto, Daisuke Miki, Kazuhiro Ishihara and Yoshiyuki Ta

Seismic Activity at Kuju Geothermal area,Kyushu,Japan

Miyuki yosinzga,Yasuaki Sudo, Tomokit Sutsui, Sachijiro Ta

Self-potential measurements on the central cones of Aso Volcano

Hideaki HASE, Yoshikazu TANAKA, Takeshi HASHIMOTO and Shin'ya SAKAI

•$B1tD>%˝%l!•$B!!J?>k5¥@WCO$NCOHW$N•(BS•$BGH$NB.EY$H•(BQ•$BCM•(B

•$B@V>>=cJ?!&@9@n!!

Gravity Basement Structure of Lijiang Basin in Yunnan Province, China

Masao KOMAZAWA, Kajuro NAKAMURA, Junpei AKAMATSU, Lequn J

Temperature Dependence of Strength of Rocks under High Confining Pressure •$B!!!]•(BMicromechanism and Implications•$B!]•(B

Mitsuhiko SHIMADA and Junk

Seismic Explosion Survey for Crustal Structure in the Shikoku and Chugoku Districts

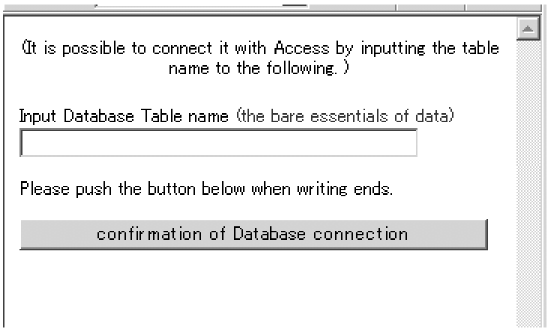
Takuo SHIBUTANI, Kiyoshi ITO, Kazuo MATSUMURA, Kunihiko WATAN
Hiroaki NEGJSHI,Yosuke FUJISAWA, Shigeyoshi TAKAHASHI, Ryohei NIS
Tatsuya NOGUCHI, Takashi YODEN, Hirotomo YOSHIKAWA and Shuhei YASU

Fig. 6 The DPRI Annuals contents

### 3.2 The HTML file of contents that was made by the PHP computer program automatically

Last year we manually made the contents file of HTML, based on the preserved literature. However, we spent a lot of time for the making the HTML files because it occasionally happened to make typos.

Consequently, we utilized the database computer program found in PHP, so that we solved the problems. In order to make the HTML file automatically, we need to instruct the PHP computer program properly, as shown in Figure 7 and Figure 8. The PHP computer program is a server-side HTML embedded scripting language. We made the HTML files automatically by the PHP computer program after pulling up the literature information preserved in the database.

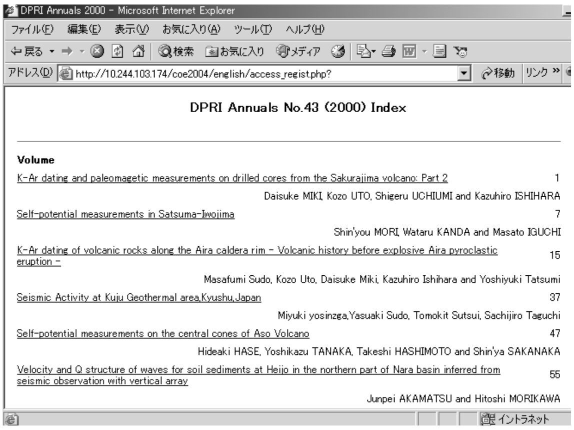


(It is possible to connect it with Access by inputting the table name to the following. )

Input Database Table name (the bare essentials of data)

Please push the button below when writing ends.

Fig. 7 PHP computer program display



DPRI Annuals 2000 - 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)
戻る 検索 お気に入り メディア
アドレス(D)
移動 リンク »
イントラネット

http://10.244.103.174/coe2004/english/access_regist.php?

DPRI Annuals No.43 (2000) Index

**Volume**

K-Ar dating and paleomagetic measurements on drilled cores from the Sakurajima volcano: Part 2 1
Daisuke MIKI, Kozo UTO, Shigeru UCHIUMI and Kazuhiro ISHIHARA

Self-potential measurements in Satsuma-Iwojima 7
Shin'you MORI, Wataru KANDA and Masato IGUCHI

K-Ar dating of volcanic rocks along the Aira caldera rim - Volcanic history before explosive Aira pyroclastic eruption - 15
Masafumi Sudo, Kozo Uto, Daisuke Miki, Kazuhiro Ishihara and Yoshiyuki Tatsumi

Seismic Activity at Kuju Geothermal area,Kyushu,Japan 37
Miyuki yosinzga,Yasuaki Sudo, Tomokit Sutsui, Sachijiro Taguchi

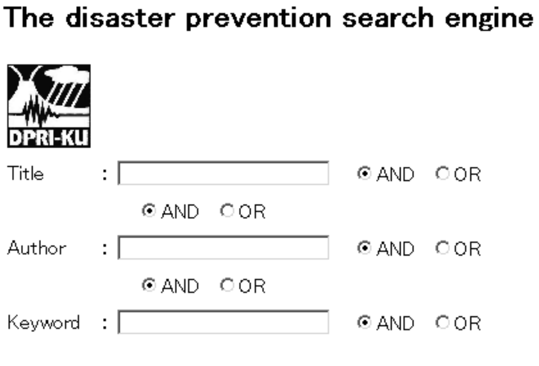
Self-potential measurements on the central cones of Aso Volcano 47
Hideaki HASE, Yoshikazu TANAKA, Takeshi HASHIMOTO and Shin'ya SAKANAKA

Velocity and Q structure of waves for soil sediments at Heijo in the northern part of Nara basin inferred from seismic observation with vertical array 55
Junpei AKAMATSU and Hitoshi MORIKAWA

Fig. 8 the HTML file that was made by PHP program

### 3.3 The new search engine that can search the disaster prevention research materials of DPRI including the DPRI Annuals

The search engine that runs on DPRI website now can only search the DPRI Annuals. It is necessary to construct a new search engine to be able to search not only the DPRI Annuals but also the other disaster prevention research materials. The new search engine can search the materials based on information from three main elements (Title, Author, and Keyword). The reason to select these three elements is that the three elements are common elements that all disaster prevention information own. We constructed the prototype of a new search engine besides three existing search engines, as shown in Figure 9.



The disaster prevention search engine

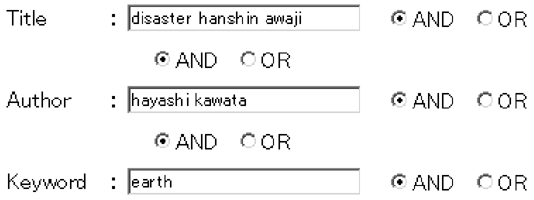
Title : ● AND ○ OR
● AND ○ OR
Author : ● AND ○ OR
● AND ○ OR
Keyword : ● AND ○ OR

Fig. 9 The new disaster prevention search engine

Main two features of the new search engine are enumerated as follows.

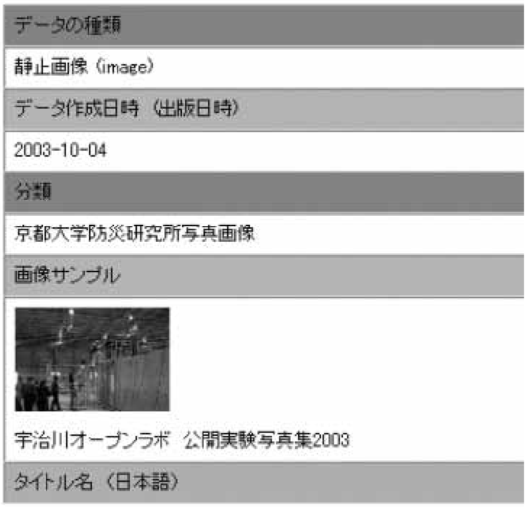
The new search engine can search the disaster prevention materials based on information from the words that are separated with a space, such as “Yahoo” and “Google”, as shown in Figure 10.

The new search engine can search information of the disaster prevention materials other than the literature, for instance static digital image, as shown in Figure11.



Title : disaster hanshin awaji ● AND ○ OR
● AND ○ OR
Author : hayashi kawata ● AND ○ OR
● AND ○ OR
Keyword : earth ● AND ○ OR

Fig. 10 the words that are separated with a space



データの種類
静止画像（image）
データ作成日時（出版日時）
2003-10-04
分類
京都大学防災研究所写真画像
画像サンプル
宇治川オープンラボ 公開実験写真集2003
タイトル名（日本語）

Fig.11 a static digital image searched

### 4. Conclusions

In this year, we have improved the work system and introduced a high-speed scanner for the increase in the work efficiency. As a result, by the end of 2004 fiscal year, we completed digitization the preserved literature in the DPRI (all DPRI Annuals, DPRI Bulletin from No.25 to No.45, DPRI extension lecture from No.1 to

No14). In addition, we constructed the prototype of a new search engine to be able to search the disaster prevention research materials of DPRI including the DPRI Annuals. The search engine can retrieve disaster prevention information other than the literature.

For the future, we are going to enrich search engine functions that offer the data of many other disaster prevention materials 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 express their sincere gratitude to the colleagues, the adjuncts 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 (in Japanese).

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

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

Hideki MATSUURA, Kenichi TATSUMI, Hideo TAGAWA, Yoshinori YOSHIDA, Tsutomu MIURA, Tetsuro TAKAYAMA, Hiroo WADA and Norio, (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, Annuals of DPRI, Kyoto Univ., No. 47 C, pp. 117-126.

Tomohiro KUGAI, Yoshiaki KAWATA and Haruo HAYASHI, (2004): Development of Cross-Media Database for Sharing Disaster Information, Annuals of DPRI, Kyoto Univ., No. 47 C, pp. 331-336.

Go URAKAWA, Nozomu YOSHITOMI*, Tomohiro KUGAI, Hironori KAWAKATA, Kenneth C. Topping and Haruo. (2004): Development of Cross-Media Database for Sharing Disaster Information and A Case Study about Implementation Process, Annuals of DPRI, Kyoto Univ., No. 47 C, pp. 337-344 .

## 印刷物・映像情報の電子ファイル化と<br>Ｗｅｂ上で高速検索可能なシステムの構築（２）

松浦秀起・辰己賢一・吉田義則・三浦勉・高山鐵朗・和田博夫・平野憲雄

### 要旨

本研究では、防災研究所に蓄積されてきた研究成果の印刷物・映像情報を電子ファイル化し、Ｗｅｂ上で高速検索可能なシステムの構築を目指している。平成16年度では、高速スキャナ、一括画像処理を取り入れた文献の電子ファイル化体制を確立し、昨年度に比べ作業効率を上昇させた。その結果、すべての年報のみならず、Bulletin、公開講座を含めた、約６万ページが、Web 上で公開可能である。また、文献以外の防災情報を高速検索できる新しい検索システムの構築案も示した。

キーワード: 防災研究所年報; OCR; 電子ファイル; 高速スキャナ; PDF; 高速検索システム