防災研年報の電子化と HP での高速検索

○松浦秀起、○辰己賢一、多河英雄、吉田義則、三浦勉、高山鉄朗、和田博夫、平野憲雄
京都大学防災研究所技術室

概要

京都大学防災研究所年報（以下、年報と表記する）は 1958 年から始まり、2003 年度には 46 巻目まで続いている。この年報を 2002 年度の第 45 号から過去に遡って 1 ページずつイメージスキャナで取り込み、1991 年度の第 34 号までの 12 年分を電子ファイル（PDF 形式）化した。また電子ファイル化した年報文献に対して、OCR（Optical Character Reader:光学式文字読取装置）処理したテキストを作成し、Web 上でも高速検索（目次検索、カテゴリ検索、全文検索）ができるようにした。

1　はじめに

京都大学防災研究所（以下、防災研究所と表記する）は昭和 26 年に災害の学理とその応用の研究を行うことを目的に設置された[1]。その後、防災研究所の研究分野は非常に多岐にわたり、「災害学理の究明と防災学の構築」の研究分野において非常に重要な役割を担っている。

今後ますます防災学研究を推進させるためには、これまで防災研究所が蓄積してきた膨大な防災資料を一般に公開し、分かり易い形で提供することが必要であり、防災研究所の基礎資料が社会にとって生きた防災情報源として付加価値を高めることになると考えた。まず手始めとして、防災研究所の基礎資料の中でも重要な位置を占める年報に注目し、年報をインターネット上で防災研究所年報を電子ファイル（PDF 形式）の形で一般公開すると同時に高速検索によって誰でも必要な情報だけを取得可能なシステムを構築してサービスを開始することを目的とした。

2　年報電子ファイル化作業概要

近年になって、情報通信技術・メディア技術の急速な発展によって、情報を扱うことが重要になってきている[2]。そして多種多様な情報がインターネット上に溢れている現代では、大量の情報を必要なときに必要なときだけ有効に利用することが必要不可欠である。防災研究所年報は、防災研究所が設立されてから 7 年経った 1958 年に第 1 号が出版され、2003 年度に出版された第 46 号まで続いている。第 46 号はすでに電子ファイル化されているため、第 45 号から過去に遡って、年報を電子ファイル化する作業を実施した。

年報は歴史が古く、ごく最近のものを除いて大半が紙の情報で保存されている。この紙の情報を電子化するためには、イメージスキャナでアナログ情報である紙からデジタル情報であるビットマップを始めとした画像ファイルに変換し、その画像ファイルから電子書類のデファクトスタンダードである PDF（Portable Document Format）書類[3]を作成した。ただし、この作業には非常に多くの人力が必要であり、2003 年 1 月～3 月までの間、非常勤職員の方を雇用した。

作業時に最も注意した点は、複雑な作業を出来る限り単純化し、流れ作業形式にすることである。これは、最小限の努力で最大の成果をあげるために必要と考えたからである。また単純化できない専門技術を必要とする複雑な画像編集作業等は、技官が行った方が無駄をなくすことができる。

流れ作業は年報の電子ファイル化で最も労力を要し、かつ最も重要である。その手順は、大きく分けて以下の 2 つである。

（1）》 イメージスキャナを使用して、年報を 1 ページずつデジタル画像ファイル（ビットマップ形式）の形で取り込み保存する。

（2）》 保存したデジタル画像は細かい汚れが目立つため、画像編集ソフトによって 1 ページずつ手動で汚れを取り除く。

（3）》 （1）、（2）の手順によって作成したデジタル画像ファイル（以下、ページ画像と表記する）を加工し、年報の文献単位で、PDF ファイルの形に変換する

以上の作業を 3 ヶ月間実施し、約 2 万ページ（12 年分）が公開可能となった。なお、日本語 OS 環境では表示できる PDF でも、英語 OS 環境では表示できない場合がある[4]ため、表示する PDF ファイルはテキストの埋め込みを一切行わず、画像のみで作成した PDF にした。

3　電子書類検索システムの構築

3.1　目次検索システム（担当責任者》 多河、松浦）

各年報の冊子ごとの目次を HTML 文書に変換する[5]。文献ごとに 2 で作成した PDF へのリンクを作成し、クリックして年報文献を表示する（図 1 参照）。以下に目次検索システムの URL を示す。

図 1.》 目次検索システム（http://www.dpri.kyoto-u.ac.jp/web_j/index_annuals.html）

3.2　カテゴリ検索システム（担当責任者》 辰己）

目次検索では表示したい文献がどの年報にあるのか知らないと、検索が不便である。そこでタイトル、著者、要旨等のカテゴリについての情報から年報の文献を検索するシステムを構築した。検索システム構成は、Apache+MySQL+PHP である。いずれもオープンソースソフトウェアであり、基本的に無料で利用できる。数多くのデータベースシステムの内、この構成を選択したのは以下の理由からである。

MySQL は Oracle、SQLServer を始めとした高額な商用のデータベースと比べ、無駄を省いた高速かつ扱いやすいフリーのデータベースである。そして、データベースと連携を取ることができるサーバスクリプト言語 PHP は、Apache のモジュールとして動作するため、別プロセスで動作する CGI に比べ OS への負担も小さく、データベースとの親和性も高い言語である。両者を組み合わせた「MySQL と PHP＋Apache」のデータベースシステム構成は、世界的に支持を受けている[6]。このようにこの構成はコストパフォーマンスに優れ、高速かつ安定といった特徴を持つため、今回のカテゴリ検索システムはこの構成を選択した。

ただし、実際に表示する年報 PDF は、画像のみで作成されている。そのため、システム構築のほか、年報 PDF の文献単位でカテゴリの情報部分に対して OCR 処理を行い、カテゴリ情報を抽出し、データベース（MySQL）に保存する作業も行った。

使用方法は、カテゴリごとに任意のキーワードを入力し、「検索」ボタンを押す（図 2 参照）。その後表示される検索結果の中から表示したい文献を探し、タイトル部分にある「pdf 閲覧」を押す（図 3 参照）ことで年報文献が表示される。

図 2. カテゴリ検索システム」（http://www.dpri.kyoto-u.ac.jp/search/category_j.tech）

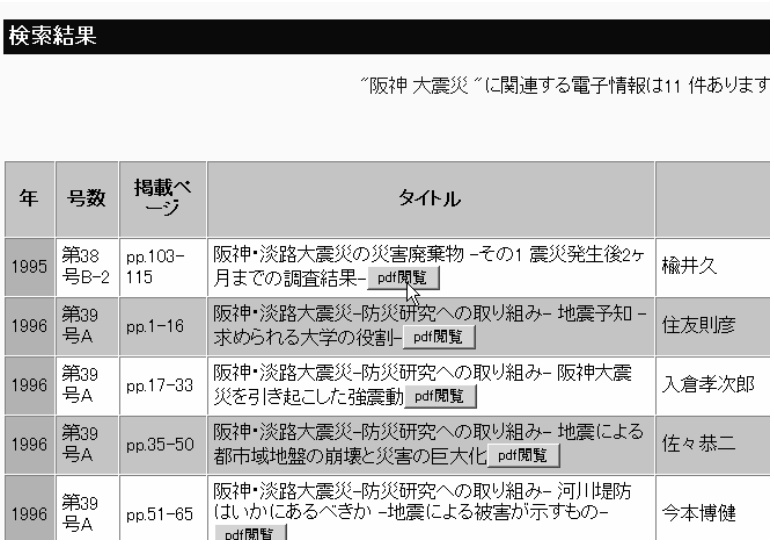

図 3.」カテゴリ検索システム実行結果

3.3　全文検索システム（担当責任者」松浦）

全文検索システムは、日本語全文検索ソフトの中でも唯一のオープンソースソフトウェアである Namazu[7] を使用した。Namazu は 1997 年、高林哲が開発を開始し、バージョン 2.0 以降からは Namazu Project によって開発が続けられている。Namazu は無料で手軽に使えることを目指した日本語全文検索システムであり、CGI として動作させることにより、小中規模の WWW 全文検索システムを構築することができるほか、コマンドラインや Emacs 上から利用するといった個人用途にも使えるシステムである。構築には、Namazu 本体に加え、CGI が使用できる Web サービス環境（Apache、Perl 等）、日本語分かち書きが使用できる環境（NKF:Network Kanji Filter、KAKASI もしくは Chasen 等）が別途必要であるが、これらはすべてオープンソースソフトウェアであるため、コストパフォーマンスは非常に良い。ただし、3.2 で述べたように、検索する対象が画像のみの PDF 文書である。そのため、テキスト形式の年報文献文書が必要である。

そこで、まず 2 の（1）で作成した文献のページごとのデジタル画像ファイルに対して OCR をかけ、画像ファイルから文字抽出を行いテキスト文書の形で保存する。そして VisualBasic6.0 で自作したプログラムによって HTML 形式に自動変換する。この HTML からさらに年報文献に対してリンクを貼るようにした。

使用方法は検索テキストボックスに任意のキーワードを入力し、Search ボタンを押す。するとキーワードが含まれている数の順に検索結果が表示されので、タイトルか、「Link is here」を押すと、年報文献が表示される（図 4 参照）。

図 4） 全文検索システム） （http://www.dpri.kyoto-u.ac.jp/cgi-bin/fullsearch.html）

4　終わりに

2003 年度防災研究所年報 Web 化は、第 34 号～第 45 号まで終了し、それら年報文献を検索できるシステムは 2003 年 5 月にテストを終えた初期バージョンが始動した。2004 年度は、引き続き残りの年報の WEB 化と年報以外の様々な防災情報を検索できるデータベースシステムの構築に取り組んでいる。

5　謝辞

今回、この年報 Web 化作業をするにあたり、協力してくださった関係者の皆様方に深く感謝致します。

参照

[1]　京都大学防災研究所, “研究所の概念”, http://www.dpri.kyoto-u.ac.jp/web_j/index_idea.html

[2]　速水 治夫,山崎 晴明,宮崎 収兄, “データベース IT Text” , 平成 14 年 9 月

[3]　総務省 行政管理局, “電子政府の総合窓口”, http://www.e-gov.go.jp/

[4]　中島 啓光, et al, “日本語環境での電子出版”, 平成 14 年度東京大学総合技術研究会報告集, 平成 15 年 3 月, P5-15 – P5-17

[5]　石橋 健一, 鐘ヶ江秀彦, “超図解 HTML タグ辞典”,） 平成 14 年 12 月

[6]　立岡佐到士, “実例で身につける MySQL×PHP による本格 Web-DB システム入門”, 平成 15 年 5 月

[7]　Namazu Project, “日本語全文検索 Namazu”, http://www.namazu.org/