防災研究所で蓄積された印刷物や映像情報の電子ファイル化とホームページで高速検索可能なシステムの構築

技術室
松浦　秀起

1. はじめに

防災研究所は昭和 26 年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ました。その当時、防災研究所は京都市左京区吉田本町にあり、第 1 部門（災害の理工学的基礎研究部門)、第 2 部門（水害防御の総合的研究部門)、第 3 部門（震害、風害など災害防御・軽減の総合的研究部門）の 3 部門に加え、宇治川水理実験所、地殻変動観測所（桜島、阿蘇、鳥取、潮岬等、２０箇所）の２附属施設、教授 11 人、助教授 8 人、助手 6 名、講師 3 名、技官 4 名だったそうです。その後、社会環境の変貌による自然災害の多様化や学問の進展とともに、社会的要求が高く新たな災害問題を研究するために、研究部門および研究センターの整備が行われました。現在では防災研究所の研究分野は多岐にわたり、5 大研究部門（19 研究分野)、5 研究センター（19 研究領域)、15 実験所・観測所を有する研究所へと発展をとげています。人員も（平成 14 年 10 月 1 日現在）教授 32 名、助教授 34 名、助手 40 名、講師 28 名、技官 26 名と大所帯になっています。平成 14 年には 21 世紀 COE プログラムの「学際・複合・新領域」分野において「災害学理の究明と防災学の構築」で研究拠点の 1 つに選ばれていて、今後期待されていると同時に注目を浴びています。

日本の中でも自然災害研究の「卓越した研究拠点」として認定された防災研究所には、年報を始め、蓄積されている論文やハザードマップなどの印刷物と実験・観測・調査など映像記録が膨大な研究基礎資料として保存されています。防災研究所年報は防災研究所が設立されてから 7 年経った昭和 32 年（当時の所長は西村英一教授）に第 1 号を出すことが決定され、それから現在発行されている最新の第 46 号まで続いています。このような年報を始めとする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は、防災学研究を推進させ、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思います。その手始めとして、防災研究所の基礎資料の中でも非常に重要な位置を占める年報をインターネット上で公開し、高速検索によって必要な情報だけを取得可能なシステムを構築することを目的としたのが、今回のプロジェクトです。技術室に研究費として数百万程度予算がついたのは珍しいということと同時に、成果を出さなければいけない非常に技術室の存亡をかけたプロジェクトでもありました。

2.　防災研究所年報の電子ファイル化及び、検索システム構築プロジェクト

2.1　プロジェクト概要

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たと思います。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す。

防災研究所も年報という形で、過去の研究成果を残してはいるものの歴史が古く、大半が紙の情報で保存されているというのが現状です。

年報だけでも膨大な量であるため、情報をどのように電子化し、そして高速かつ簡易に検索するシステムをどのようにして構築するかは、議論されるべき問題でした。まず人力が非常に必要な作業は、出来る限り単純化し、最小限の努力で最大の成果をあげるようにすることが必要です。これにより予算の半分以上が、人件費に使用することが決定され、実際には非常勤職員を１月に５人雇用し、２月に１人追加で雇用し、合計１７人月の労力を補助として使用できました。

ただ問題が一つありました。それは非常勤職員の方々のほとんどがパソコンに不慣れということです。そのため、複雑な作業は止め、年報の電子ファイル化で最も労力を要し、重要な以下の二点の手順を中心に行うことにしました。

＜1＞ イメージスキャナを使用してアナログの紙の情報を一ページずつ読み込んで、デジタル画像ファイル（ビットマップ形式）の形で保存する。

＜2＞ 保存したデジタル画像に対して OCR（Option Character Reader:光学式文字読取装置）を使用することによって画像から文字を抽出し、テキストファイルの形で保存する。

＜1＞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）と＜2＞の手順によって作成したテキストファイル（以下、OCR テキストと表記する）の二つのファイルがあれば、後は加工次第で良いものが作成できると考えていました。最終的には年報の論文単位で、電子書類のデファクトスタンダードである PDF（Portable Document Format）ファイルの形に変換して提供し、その電子書類の検索する手段として、以下の三つの検索システムを構築しました。

(A) Web サイトで目次を表示して、そこで文献の検索を行う「目次検索」

(B) タイトル、著者等の情報を年報の文献データベースに問い合わせた結果を元に、文献の検索を行う「カテゴリ検索」

*「簡易検索」も別途作成したが、これはカテゴリ数を減らした「カテゴリ検索」です。

(C) 好きなキーワードを入力し、そのキーワードについて年報の文章自体をすべて検索した結果を元に、文献の検索を行う「全文検索」

本プロジェクトでの主な作業項目とその内容について以下に列記します。

(1) プロジェクトの企画、立案

(2) 年報の電子ファイル化のための準備

(3) 年報の電子ファイル化作業の実施及び、ページ画像と OCR テキスト等電子ファイルの管理

(4) ページ画像から PDF ファイル形式へのファイル形式変換作業

(5) 電子書類検索システムの構築

(6) 電子ファイル化された年報文献と検索システムのチェック及び公開

2.2　プロジェクト詳細

(1) プロジェクトの企画、立案

プロジェクトの企画が持ち上がったのは、10 月前後でした。本格的に始動するのは、1. 月からであったが、それまでに決めておかなければいけない事項がありました。具体的に決める必要があったのは、主に作業全体の流れを整理することと、(3)、(4) 以外の作業分担、年報データ管理の手法の確立です。

つまり全体の作業の 80%以上を占める(3)、(4)の作業を非常勤職員の方の労力にあて、それ以外の (1)、(2)、(5)、(6) は専門的知識かつ作業が複雑化するためこれらは技術室が行う必要があると考えました。実際に全体の作業を単純化したリストを、作業名、作業概要、作業担当者（＊は作業責任者）の順に以下に示します。

・広報活動

予算管理、発表等：本プロジェクトの広報、発表

＊平野・多河・吉田・松浦

・ 年報収集

年報原本の収集

＊三浦・多河・高山・和田

・ プロジェクトの企画、監督

プロジェクト自体の作業の流れを企画し、作業が始まるにあたっての前段階の設定や必要な情報等の収集、作業全体の監督

＊松浦、平野、多河、吉田、三浦、辰己

・ 年報電子化

紙の年報をスキャナによって電子ファイル化（OCR や年報の主要項目（発行年、題名、著作者、概要等）のデータベース化作業を含む）する作業、及び年報電子ファイルの管理、ライブラリアン（非常勤職員）の教育と監督

＊松浦・西村・多河・吉田

・ Web システム、コンテンツデザイン構築

年報や検索システムを搭載する Web システムやコンテンツデザインの構築

＊辰己・吉田・西村・松浦

・ 年報検索システム構築１（目次検索）

目次が表示される HTML から年報を検索するシステムの構築

＊多河、松浦

・ 年報検索システム構築２（SQL 検索）

年報の主要な項目をデータベース化し、そこより検索するシステムの構築

＊辰己

・年報検索システム構築３（全文検索）

年報文章全体から検索するシステムの構築

＊松浦

年報のデータ管理は、ページ画像と OCR テキストを主に対象としました。後でばらばらになってもすぐに整理して編集できるように、ファイル名で分類をすることを考えまし

た。ファイル名の規則を会議で話し合い、作成されたページ画像がどの年報のどこのページにあるかが分かるように決定しました。以下に詳細を示します。

(例)　a00200045b2015021b01　2000 年度第 45 号 B2 の 15p～21p の論文の 1 ページ目

左 1　　書類（年報が対象なので今回はすべて a）
左 2、3　　予備桁（通常は 00）
左 4～7　　年報の作成年度
左 8、9　　号数
左 10、11　　年報の種類（A なら a0、B1 なら b1、B2 なら b2）
左 12～14　　イメージスキャンする論文の最初のページ
左 15～17　　イメージスキャンする論文の最後のページ
左 18　　ファイルの種類（b:ページ画像、t:OCR テキスト）
左 19、20　　イメージスキャンするページ画像が論文の何ページ目かを示します

（2）年報電子ファイル化のための準備

（3)、(4）の作業に向けて、作業環境を整える必要があります。以下に（3)、(4）の作業において使用したものを示します。

＜主要機器＞

PC、スキャナ（GT9300－UF）
大容量 HD（120GB）

＜使用ソフト＞

画像編集ソフト：Irfan（フリーソフト)、Paintshop7.0
OCR ソフト：WinReader8.0、読ん de!!ココ 3.0
その他：Microsoft Access2000、VisualBasic 6.0

（3）年報の電子ファイル化作業の実施及び、ページ画像と OCR テキスト等電子ファイルの管理

1 月より非常勤職員 5 名に 1 日間作業に関する教育を実施した後、D-170 号室にてイメージスキャン作業を開始しました。読ん de!!ココのソフトを使用し、自動実行で作業をできる限り単純化し、作業手順やファイル名の規則性などは、パワーポインタで作成したポスターをすべての作業者が随時確認できるように壁に貼りつける他、出来る限り非常勤職員の方を即座に指導、補佐できるようにしました。

一番大変だったのは、ページ画像のチェックでした。1 日平均 700 枚程度ですが、非常勤職員が帰宅後、すべてのページ画像に対して間違いがないかどうかのチェックを行い、間違いがあれば次の日に訂正するようにします。間違いは一日平均 10 枚程度ですが、間違いがあっては、後で修正が困難なので、ページ画像については出来る限り重複してチェックを行うようにしました。こうして作成したページ画像は、(4）の作業に入る前にすべて Paintshop7.0 を使用してゴミを取ります。

実際の収集してきた年報は保存状態が良いものが多かったのですが、スキャンしてみると細かい汚れが目立つものが大半でした。これについては、どこまで拡大してゴミをとるかということを考えましたが、結論としては A4 印刷して印刷されない程度の汚れは取らない方針でいきました。やはりこの作業は時間がかかるため、大きな汚れ以外の除去については、時間を費やしても成果は上がらないと考えたからです。他に（3）の作業実施の

際に気をつけた点を以下に示しておきます。

＜注意点＞

・ チェックシートを作成し、ページ画像一枚ごとにチェックシートに作業者が書き込む
・ 図、表等、スキャンしても見難いページ、カラーページは、Paintshop7.0 のソフトによって、カラー&高解像度でイメージスキャンし、ページ画像を再度作成し直す。
・ すべてのページ画像と OCR テキストは、大容量 HD に毎日バックアップを取り、不測の事態に対処できるようにした。

次に（3）の作業実施時に問題となった点とその対応策を示します。

＜問題点＞

・ 年報の素材源そのものが歪んでいることがある。

→部分的に歪んでいるものについては、補正すると全体が歪むため、補正せずにそのままにしました。

・ ページ画像のファイルの種類であるビットマップ画像や、カラーで取り込んだ画像はファイルサイズが大きい。

→Irfan ですべてのページ画像の画質、解像度を落として統一し、カラーで取り込んだ画像に関しては、それに加えて減色しました。

・ ページ画像すべてに汚れが目立つ

→Paintshop7.0 によって、すべてのページ画像に対してゴミ取りを行いました。なるべく、一枚のページ画像あたりのゴミ取りは五分を超えないようにします。理由は、あまりにも小さいゴミや、見つけにくいゴミもありますが、これは拡大して初めて見つかるゴミであるから、目立つゴミを優先して削除していくように指導しました。

（4）ページ画像から PDF ファイル形式へのファイル形式変換作業

この作業は、WinReader8.0 を使用し、論文単位でページ画像から画像 PDF ファイルを作成しました。実際には、Irfan によってサイズを落とした後のページ画像を PDF ファイルに変換していますので、文字の潰れ等に注意して作成する必要があります。

（5）電子書類検索システムの構築

一般公開する（4）で作成した PDF は画像のみのファイルです。これは、Acrobat の英語版では、日本語フォントが入っているとエラーが出て、文章が表示できないことが判明したからです。画像のみの PDF ファイルなら、どのような Acrobat であっても文章が画像で表示できるためこのようにしました。しかしこのままでは検索できません。そこで、OCR テキストを元に検索システムの情報元をデータベース化する必要がありました。この作業には、非常勤職員の労力 17 人月の内、3 人月を使用しました。データベース化には Microsoft Access を使用しました。Excel でもよかったのですが、OCR テキストからコピーする文章が長いとセルからはみ出して、うまくいかない場合があったからです。検索システムで私が担当したのは、目次検索の一部と全文検索です。以下に概要を示します。

＜目次検索＞

年報の目次を HTML 文章化し、それぞれの論文タイトルには、PDF ファイルへのリンクをつける。使用方法は以下の通りです。

まず、DPRI の左のメニューより、「防災研究所年報」をクリックし、図 1 のような画面を出します。そして、閲覧したい年報文献を年報発行一覧より選んで、クリックします。

図 1 では、第 45 号の A をクリックしています。

図 1　年報発行一覧ホームページ

クリックすると、図 2 のように、HTML 文章化された目次が表示されるため、そこで閲覧したい文献を選んでクリックすると、図 3 のように PDF ファイルが表示されます。

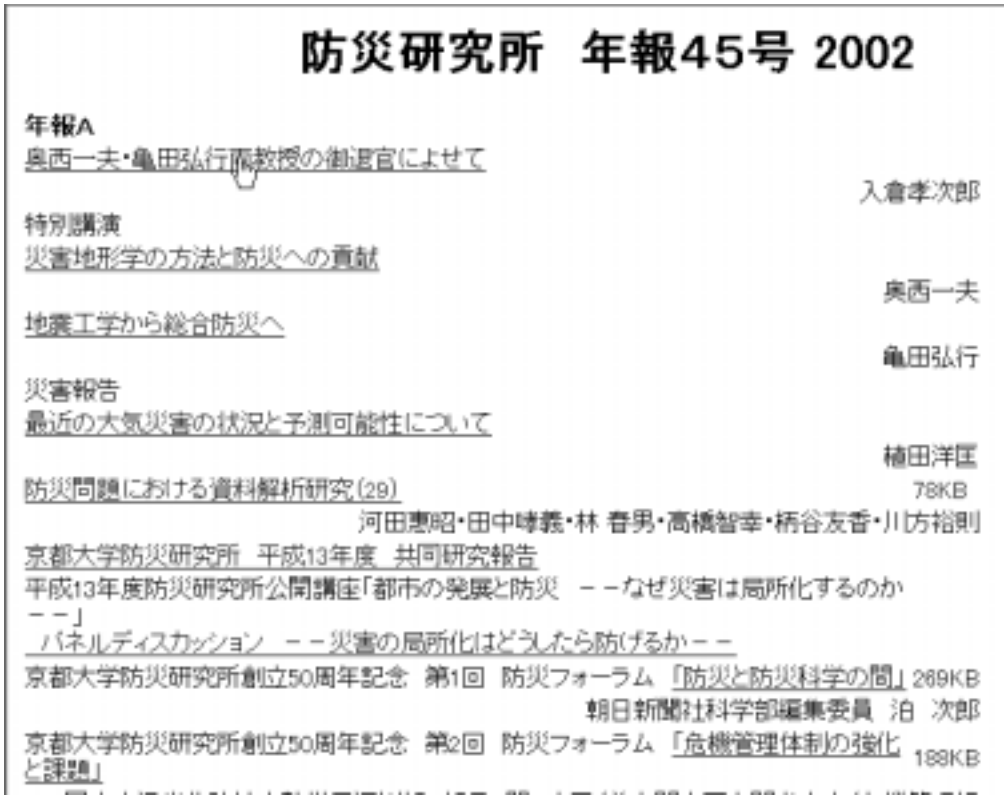

図 2　目次検索ページ

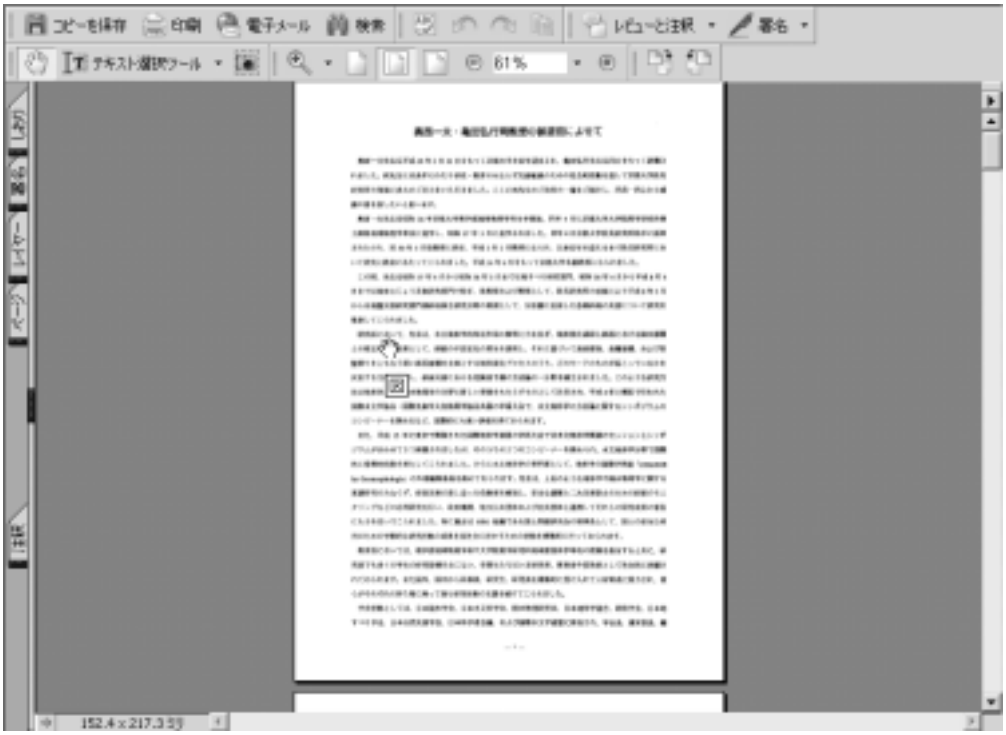

図 3　年報 PDF ページ

＜カテゴリ検索＞

システム構成は、PHP＋MySQL の検索システムです。（詳細は辰己氏に）

カテゴリ検索をクリックすると、図 4 のようなページになります。ここに、色々なキーワードを文字入力して検索ボタンを押します。

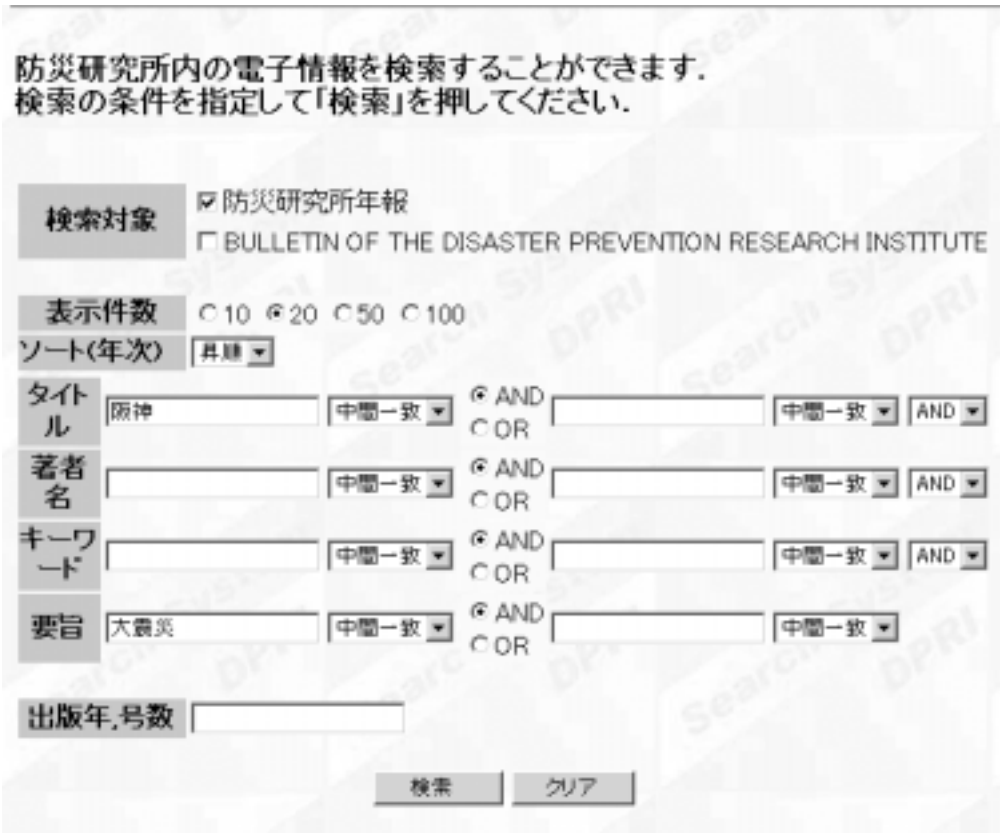

図 4　カテゴリ検索

すると下のように検索結果がでてきますので、図 5 のように pdf 閲覧のボタンを押すと、図 3 のように PDF 文書が表示されます。

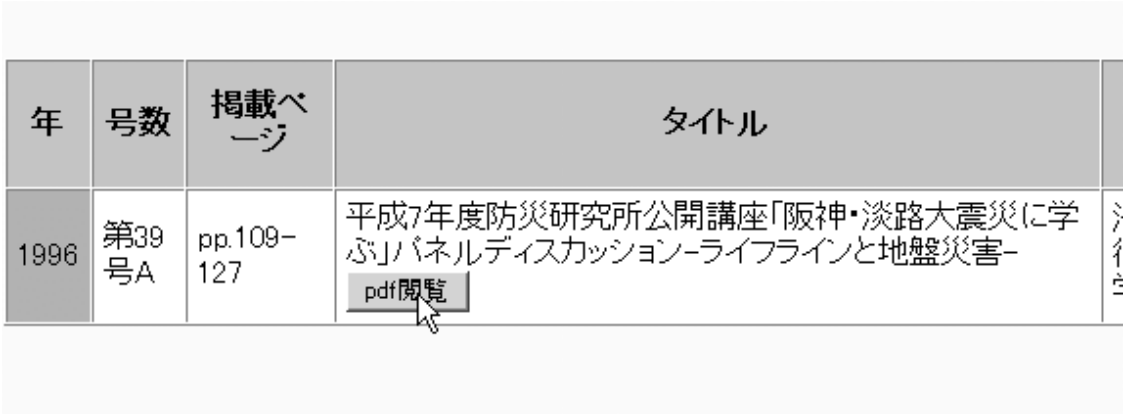

図 5　カテゴリ検索の検索結果

＜全文検索＞

主に、日本語全文検索ソフト Namazu と日本語分かち書きソフト kakasi を使用します。両方ともフリーソフトです。まず OCR テキストを元に HTML 文章を作成します。手書きでやると 1 ヶ月以上かかりそうな作業なので、VisualBasic6.0 で変換作業プログラムを作成し、自動変換します。これにインデックスを作成し、検索できるようにします。

全文検索の使用方法はいたって簡単で、Yahoo、Google を始めとした Web 検索システムと全く同じように検索できます。図 6 のように調査したいキーワードをテキストボックスの中に書き込み Serch ボタンを押します。

図 6　全文検索

すると、図 7 のようにそれに関連した文献の一覧がでてきますので、そこで閲覧したい文献のタイトルをクリックすると図 3 のように PDF 文書が表示されます。

図 7　全文検索結果一覧

3.　使用ソフトウェア紹介

＜Paintshop＞

画像編集商用ソフトです。非常に使いやすく画像編集の商用ソフトとしては、優秀だと思います。画像を編集するだけでなく、イメージスキャナのインターフェイスソフトとしても使用できます。パワーポイントのように図形描画も可能であり、今後もメインで使っていきたいソフトウェアです。

＜Irfan＞

画像ビューアのフリーソフトです。画像ビューアとしての位置付けのフリーソフトですが、実際はガンマ補正や明るさ調整、画像のリサイズ、形式一括変換等、多数の機能を備えた画像編集ソフトでもあります。操作性にも優れています。オリジナルはオーストリアの Irfan Skiljan 氏が作成し、現在こちらのページ（http://www8.plala.or.jp/kusutaku/）で日本語版が公開されています。

＜読ん de!!ココ＞

代表的な有名商用 OCR ソフトです。今回使用したのは、Virsion3 のパーソナル版で PC におまけで付いてきた廉価版です。現在の最新バージョンは Virsion9 です。そのため活字認識はそれほどよくはなく、カラーには対応していません。しかし、動作は問題なく、簡単操作で高速に動作するため、今回最もメインに使用したソフトです。

＜WinReader＞

最高峰の認識率･ネットワーク対応･大量処理･様々なドキュメントの管理/活用します。他のソフトとの違いというのは、旧漢字を多目的原稿も高精度に認識可能であり、英語認識は世界最高峰らしいです。非常に多機能で多彩な保存形式をもっているため、認識結果を PDF 形式で出力する機能に加え、画像にテキストを貼り付ける（テキスト埋め込み型 PDF）も作成可能です。

＜MySQL＞

データベースサーバのフリーソフトと言えば、現在は MySQL と PostgreSQL です。両者とも RDBMS（Relational DataBase Management System：関係データベース管理システム）であり、商用データベースと比べてもひけをとらない性能をもっています。少し前までは、PostgreSQL が主流だったのですが、ここ 2、3 年は、MySQL の方が人気も高いです。ほとんどのデータベースサービスは Web サービスと連携を取っています。連携の仲立ちをするものとしては PerlCGI が主流でしたが､今は PHP が主流になりつつあります。

＜Namazu＞

Namazu は手軽に使えることを第一に目指した日本語全文検索システムです。CGI として動作させることにより小中規模の全文検索システムを構築できます。様々なバージョンアップを得て、現在に至っています。非常に歴史があり、オープンソースであることから動作が安定しています。著作権は野首貴嗣さんにあり、http://www.namazu.org/からダウンロードできます。

＜VisualBasic＞

Windows のソフトを自作できるソフトとして､Microsoft 社が制作して販売しています。VisualBasic6.0 を境に大幅なバージョンアップが行われ、現在の最新バージョンは VisualBasic.NET2003 です。今回のプロジェクトで使用したソフトは VisualBasic6.0 で、まだ.NET バージョンより使い勝手はいいと思います。簡単なプログラムを自作するのに適しており、小回りが聞くソフトウェアです。

4. プロジェクトの結果報告

COE 研究は本来長い研究期間を想定して計画を立てていたのですが、諸事情により 2003 年 3 月で一旦区切りをつけ成果を出さなければいけませんでした。そのため作業の最終目標が平成分（第 33 号～第 45 号）までと方向転換する必要がありましたが、検索システムのプロトタイプも 3 月に行われた発表会で公開できるまでに進み、成果として報告できました。2004 年 3 月には、第 32 号～第 30 号を目安に現在進行中です。

5. 終わりに

技術室に勤めて、初めて大きな仕事を任され、本当の所は不安もありました。しかし、その不安も忙しさのあまり吹き飛びました。確かに大変な作業でしたが、今回のプロジェクトの成果は、技術室の皆様と力を合わせてがんばった結果だと私は思っています。

技術室に COE 研究予算を分配してくださった当時の防災研究所所長の入倉先生、このプロジェクトを任せ、経験の浅い私を支えてくださった平野室長、協力してプロジェクトを進め、アドバイスやサポートしてくださった多河様、吉田様、三浦様、辰己様、高山様、和田様、西村様、非常勤職員 6 名の皆様方に深く感謝致します。