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（２）

技術室
松浦　秀起

1.　はじめに

京都大学防災研究所は創立以来、わが国における自然災害研究の中心的役割を担うと共に、全国６地区の災害資料センターの中央センターとして機能し、また全国にまたがる自然災害科学総合研究班の総括的業務を長年にわたって実施するなど、災害科学研究者による研究ネットワーク形成を主導してきた。そして研究の一環として年報を始め、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きた。このような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考えられる。
よって本プロジェクトはこれら防災資料をインターネット上で公開し、ホームページ上での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。平成１４年１月から三ヵ年計画で始まった本プロジェクトであるが､多くの協力者のもと、昨年度までで年報第３２号までの電子ファイル化とそれら年報がホームページ上で高速検索できるシステムが完成した。

しかし、防災研究所が蓄積してきた膨大な防災資料は年報だけではなく、ブリテン、公開講座、静止画像、動画像等も含まれる。つまりその防災資料をインターネット上で公開し、いつでも必要な情報を高速検索によって提供できる防災情報システムの構築と、今後増えつづける防災資料の収集、電子ファイル化、管理手法の確立が本プロジェクトの最終目標であると考えられる。ただしその目標を達成するには、二つの問題点をクリアする必要があった。

一つは防災資料の電子ファイル化作業にかかるコストを前年度より少なくとも半分以下に削減しないと年報すべての電子ファイル化ですら不可能であることである。つまり平成１４年の実績で、電子ファイル化にかかる人手と時間を計算すると年報すべての電子ファイル化は今年度で終了しないということが予想された。二つ目は、現状の検索システムでは、限られた定型のドキュメント資料しか検索できないため、拡張性に乏しいということである。本稿では、これらの問題点の解決方法及び、本プロジェクトの総括を報告する。

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　プロジェクトの概要

情報通信技術・メディア技術の急速な発展によって、情報を大量に扱うことが容易になってきた。そしてその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重

要である。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が、非常に歴史が古く、最近までは紙の情報だけしか残っていないというのが現状であった。年報だけでも膨大な量であるが、今後、過去の紙情報をどのように電子化し、そして電子化された情報を高速かつ簡易に検索するシステムをどのように構築するかというのは、重要である。

２.２　昨年度までのプロジェクト進行状況

集中的に作業を行った平成１５年１月～３月までは、合計１７人月（注：１人が１ヶ月間働いたときの労力を１人月とする。つまり、５人が３ヶ月、１人が２ヶ月働いた場合は、５×３＋１×２＝１７人月の労力となる。なお、これには技術員、研究支援研究員の労力は考慮していない。）費やして、年報第４５号～３２号までの１４年分の電子ファイル化及びデータベース化が終了した。また、その後も約１年半の間、処理を続けたが大きな予算がつかないため、平成１６年８月までで３１号～２５号までの７年分、合計２１年分が完了するに留まった。

２.３　今年度のプロジェクト計画

今年度は、平成１６年９月～平成１７年３月までの期間、８人月＋追加予算で６人月の合計１４人月が使用可能であった。しかし従来と同様のやり方（１４年分の作業を１７人月の労力で行う）では、残り２４年分の電子ファイル化作業を行うことはできず、年報すべての電子ファイル化及びデータベース化は困難であることは明白であった。
そこで平野室長の提案でスキャニングの短縮化に取り組んだ結果辿り着いたのが「ドキュメントスキャナ」による高速スキャニングである。価格と機能から選択したドキュメントスキャナは、富士通　FI-5110EOX2 [USB 接続]であった。
従来のスキャナによるスキャニングは、１ページあたり平均で６０秒程度であった。それに対しドキュメントスキャナによるスキャニングは、解像度にもよるが、両面１枚あたり１２秒～４秒である。今回は３００dpi でスキャニングを行ったため、１分間に１０ページのスキャニングであった。原稿紙の入れ替え、紙詰まり（ある一定の割合で発生するため、常に監視する必要がある。）等のロス時間を加味しても、従来の約８倍のスキャン効率が見込めた。ただし、スキャン後の画像処理は従来に比べて約３倍の処理を必要するため、結果的に電子ファイル化までのコスト（本稿でのコストとは、人的コストであり、労力及び作業時間を指す）が３分の１まで削減できた。
従来の電子ファイル化作業効率は、１人月で約１年分弱であり、残り２０年分だと最低約２０人月必要であるが、「ドキュメントスキャナ」の導入によって、従来の労力の３分の１にあたる約７人月あれば、年報すべての電子ファイル化は可能であると計算できる。
また電子ファイル化のほかには、スキャニングした資料のデータベース化に４人月（１０年分あたり約２人月）と、Ｗｅｂでの公開用ＨＴＭＬファイル作成１人月が必要である。これに関しても詳細は後ほど述べるが、MS-Access の機能をうまく使えば、約４０％のコスト減少させることに成功し、３人月のコストで作業は終了した。
年報の見通しがたったので、ブリテン、公開講座の電子ファイル化も計画した。ブリテン

は第２０号～第４５号、公開講座は第１回～１４回の電子ファイル化を予定した。これらを年報に換算すると、約９年分あり、それに要する作業労力は９÷３の３人月と、追加のデータベース化とＷｅｂでの公開用ＨＴＭＬファイル作成で２人月は必要となるため、合計作業労力は５人月と予測できた。

以上、最低限今年度に必要な電子ファイル化作業の必要労力は、多く見積もっても１５人月以下（技術員、研究支援推進員の労力は、これだけにかかりきりでないため省く）であり、技術員、研究支援推進員の労力をプラスすれば。今回の電子ファイル化の手順は以下の通りである。

３．　電子ファイル化作業の詳細

３.１　スキャニング作業

ドキュメントスキャナを使用してアナログの紙の情報を読み込んで、ページごとにデジタル画像ファイル（JPEG 形式）の形で保存する。JPEG 形式になったのは、ドキュメントスキャナの仕様（表 1 参照）のためである。今回は「カラー・３００dpi」に相当する「スーパーファインモード」でスキャニングを行った。モノクロでもスキャニング可能であったが、モノクロでスキャニングを行うと取り込んだ画像の質が著しく低下するため、後で適切な画像処理を行い、モノクロに変換するときれいになるため、カラーで取り込んだ。

表１　ScanSnap fi-5110EOX2 の仕様

製品名		ScanSnap fi-5110EOX2
読取方式		自動給紙方式 (オートドキュメントフィーダ) 両面同時読み取り
読取モード		カラー / 白黒 / 自動 （カラー・白黒の自動識別)
光学系/光源		非球面レンズ縮小光学系 CCD 採用 / 白色冷陰極管
読取速度 [注：(A4 縦の例) 環境によって異なる場合がある。 PDF､JPEG ファイルで保存可能]	ノーマルモード	カラー150dpi、白黒 300dpi 相当: 両面・片面 15 枚/分
	ファインモード	カラー200dpi、白黒 400dpi 相当: 両面・片面 10 枚/分
	スーパーファインモード	カラー300dpi、白黒 600dpi 相当: 両面・片面 5 枚/分
	エクセレントモード	カラー600dpi、白黒 1,200dpi 相当: 両面・片面 0.5 枚/分

３．２　画像処理

カラーで取り込んだページごとのデジタル画像（以下、ページ画像と表記する）は、明るさ、コントラストが非常に薄く、黄ばみがかっているため、PDF 化作業（取り込んだページ画像を論文単位でまとめて PDF で保存しなおす作業）を行うと、非常に汚く、文字も判別不可能の状態であった。それを解決するため、明るさ、コントラスト、ガンマ値の補正を行い、必要があればゴミ取りの画像処理を行った。(図１参照：　表２の周囲切取については、３.３で詳細に述べる)

ページ画像の明るさ等のある程度の画像補正は、効率化のため一括画像処理ソフト「Irfan」を使用するが、ページ内に文字だけでなく「図・表」が存在するページ画像は、一括変換では「図・表」が非常に黒っぽくなり過ぎて閲覧不可になった。そのため、ページ内に「図・表」が存在するページ画像に関しては、画像以外の部分の文字部分のみをより濃くして、画像は適度に手動調整し、ゴミも手作業で取り除くことを実施した。

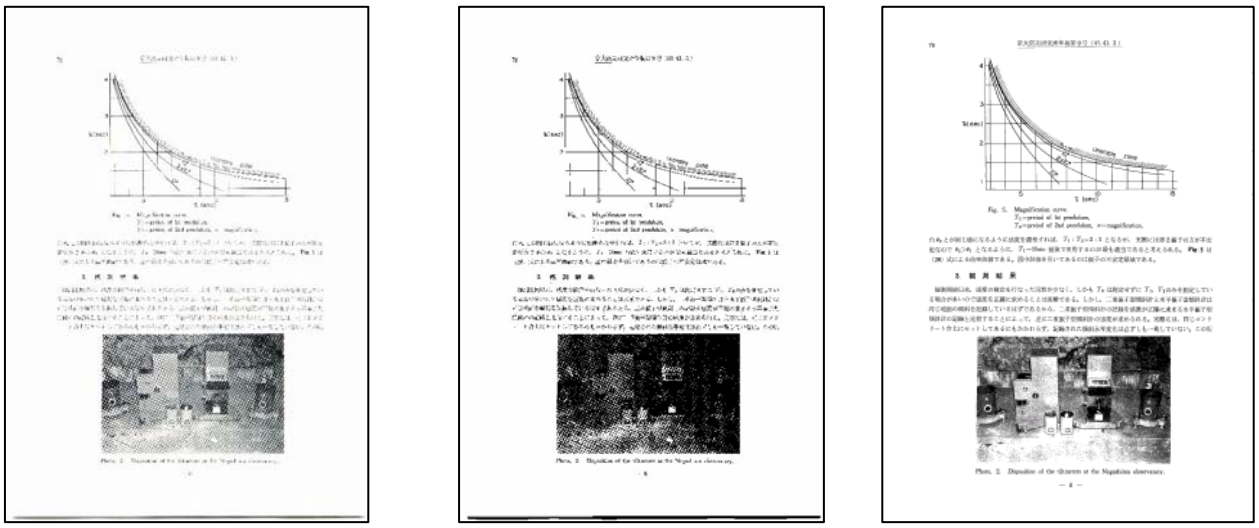


図１　画像補正
（左から　スキャニング直後、自動画像補正後、手動画像補正後のページ画像）

３．３　PDF 化作業

２で保存したページ画像を論文単位でまとめて WinReader8.0（OCR 商用ソフト）を使用して PDF で保存しなおす作業（PDF 化作業）を経て Web 上にアップする。この作業は前年度とほぼ同じ作業であるが､作業効率や品質を上げるため工夫した作業は､二つある。「外周部分の汚れを取り除く作業」､「傾き、位置補正」である。これは、冊子という紙媒体をスキャンしてページ画像にするため､外周部分に黒い線が汚れとなって表示されたり、スキャン時に傾いたままでスキャニングされたり、スキャン位置が適正でないものが出て

くる。また、原稿元自体がすでに傾いて、文書の印刷位置が適正でないものが多数あった。前年度は、これを非常勤職員雇用期間終了後も手作業で修正していたが、この作業だけでかなりの時間を要するため、今年度は、「外周部分の汚れを取り除く作業」を、図２のように外周部分を汚れがある部分を自動で一括して切り取る（使用したソフトは、trim Version １.３ / 画像トリミングフリーソフト）ことによってコストを削減した。

「傾き、位置補正」については、自動化できなかったため Paintshop（画像処理商用ソフト）を使用し手作業で修正を行った。その際に、１つ１つ傾きをチェックしては修正ではなく、他のすべてのページ画像修正が終了した時点で、順番等の最終チェックと同時に一斉にチェックを行い、修正するときには、無駄がないように「傾き、位置補正」の修正のみを行うようにした。これによってかなりのコスト削減につながったと考えられる。

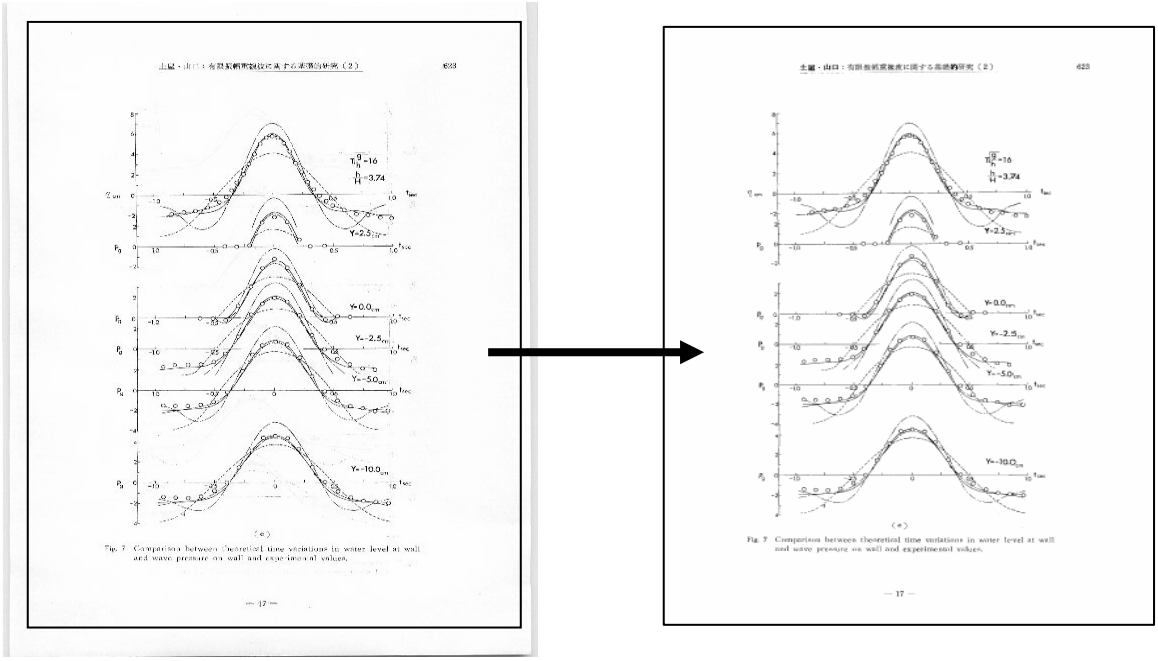


図２　外周部分の汚れの除去

３.４　データベース化・文献閲覧用目次 HTML 作成作業

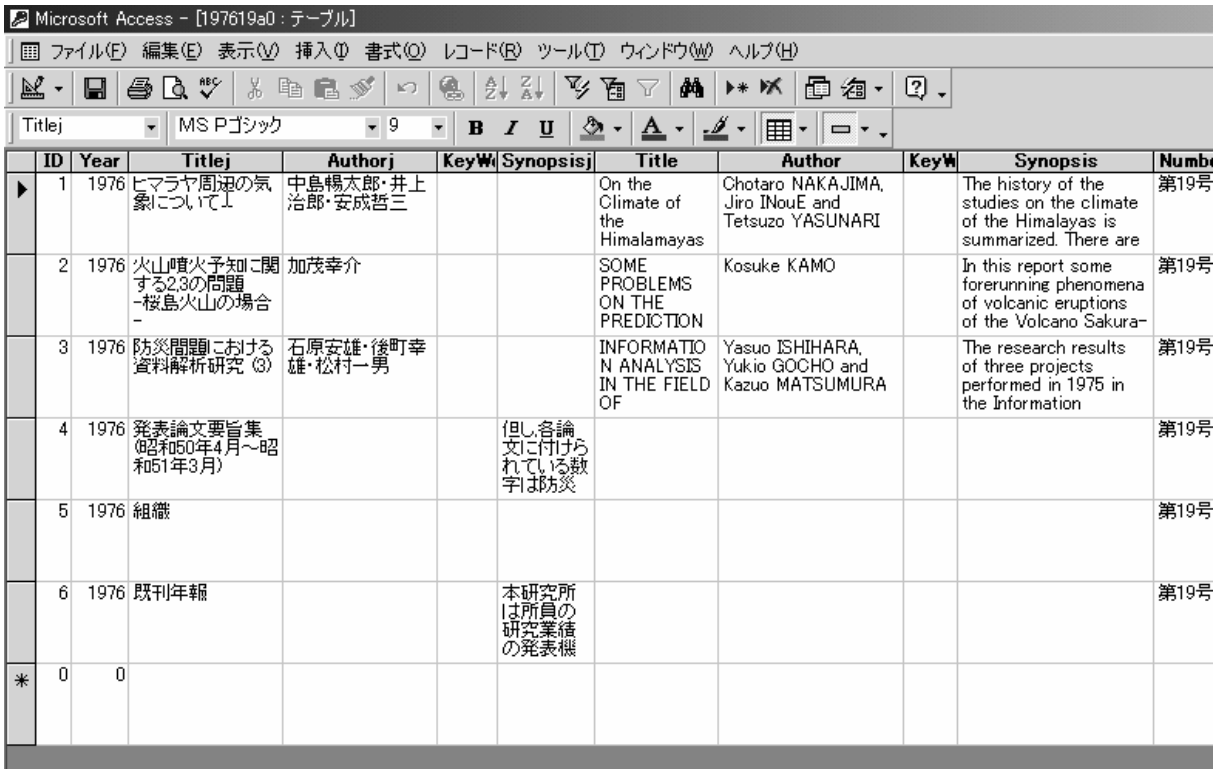
２の作業が終了したページ画像に対して、OCR（Option Character Reader:光学式文字読取装置）を使用し、ページ画像から文字を抽出し、テキストファイルの形で保存する。OCR での文字抽出時の文字正答確率は９割程度だが、このテキストファイルを使用した現在稼動中の Namazu 全文検索システムは実用に耐えるものと判断したため、前年度と同じく全文検索システムに組み込んだ。

さらに、各論文ごとの「タイトル」、「著者名」、「出版年・号数」、「要旨」部分は、MS-Access ソフトを使用して正確にデータベース化する作業を実施した。

MS-Access へのデータ入力は、非常勤職員の方にして頂いた。前年度は、図３のようなテーブル入力を採用していたが、後のデータチェックにおいて多数のデータ入力ミスが発見された。

理由として、図３のようなテーブルでの直接入力は、入力後に入力データのチェックが非常にし難いためと考えられた。そのため、今年度は、図４のようにフォームを作成し、フォーム入力を導入することによって、データ入力ミスの削減と作業効率を高めるよう工夫した。データベース入力作業終了後は、目次検索に使用する「文献閲覧用目次 HTML」（図５参照）を作成した。前年度は、HTML ファイルはデータベースからコピー&ペーストして、手作業で作成していたが、これも後のデータチェックにおいて多数の間違いが発見された。

そのため、今年度はオープンソフトウェアの PHP（Hypertext Preprocessor）を使用することで Access からのデータを抽出し、目次 HTML を完全ではないが、ある程度まで自動作成する補助プログラムを自作し、データベースからのデータコピーミスを削減し、作業効率を高める工夫をした。さらにできた目次 HTML を Web 上にアップし、PDF ファイルと見比べてタイトル、著者のチェックをすることで、データベース部分の入力ミスの再チェックを行うことも出来たため、非常に効率のよい方法であったと考えられる。


Microsoft Access - [197619a0 : テーブル]

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) レコード(R) ツール(T) ウィンドウ(W) ヘルプ(H)

Titlej | MS Pゴシック | 9 | B I U

ID	Year	Titlej	Authorj	KeyW	Synopsisj	Title	Author	KeyW	Synopsis	Numb
1	1976	ヒマラヤ周辺の気象について	中島暢太郎・井上治郎・安成哲三			On the Climate of the Himalamayas	Chotaro NAKAJIMA, Jiro INouE and Tetsuzo YASUNARI		The history of the studies on the climate of the Himalayas is summarized. There are	第19号
2	1976	火山噴火予知に関する2,3の問題 -桜島火山の場合-	加茂幸介			SOME PROBLEMS ON THE PREDICTION	Kosuke KAMO		In this report some forerunning phenomena of volcanic eruptions of the Volcano Sakura-	第19号
3	1976	防災問題における資料解析研究 (3)	石原安雄・後町幸雄・松村一男			INFORMATION ANALYSIS IN THE FIELD OF	Yasuo ISHIHARA, Yukio GOCHO and Kazuo MATSUMURA		The research results of three projects performed in 1975 in the Information	第19号
4	1976	発表論文要旨集 (昭和50年4月～昭和51年3月)			但し,各論文に付けられている数字は防災					第19号
5	1976	組織								第19号
6	1976	既刊年報			本研究所は所員の研究業績の発表機					第19号
0	0									

図３　Access テーブルでのデータベース入力

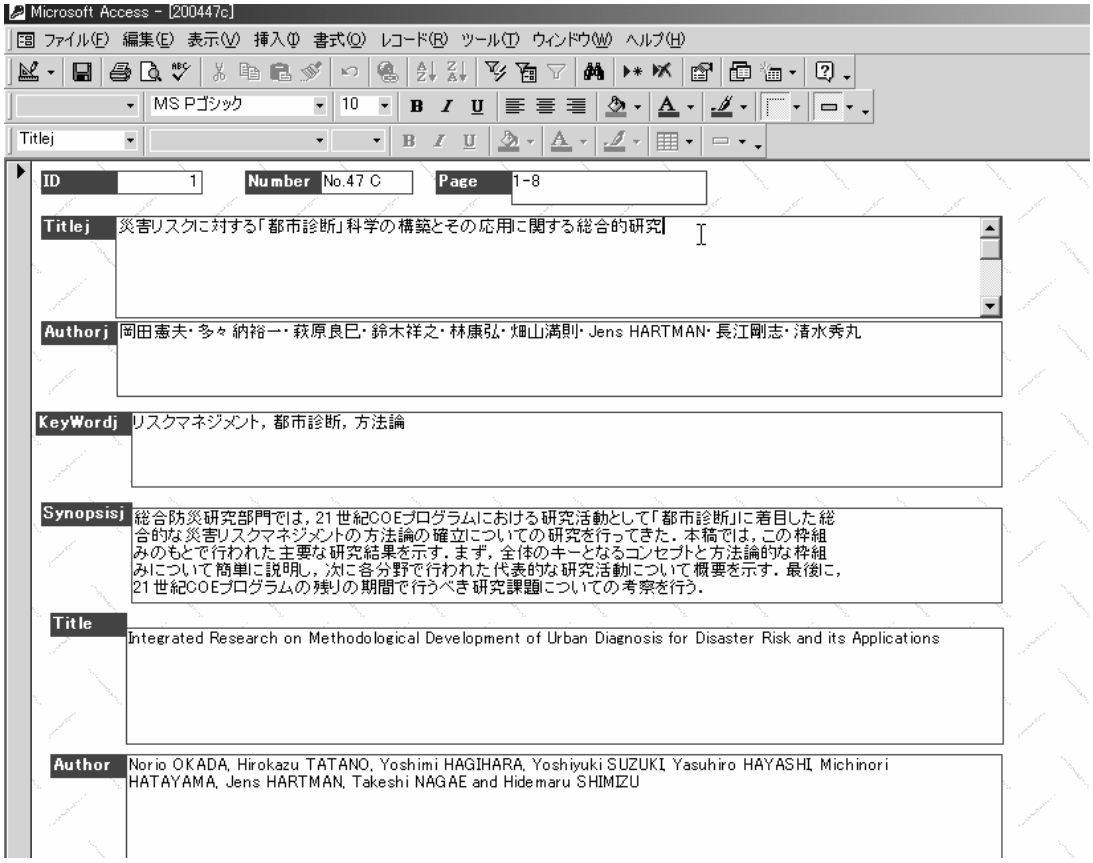


図４　Access フォームでのデータベース入力画面

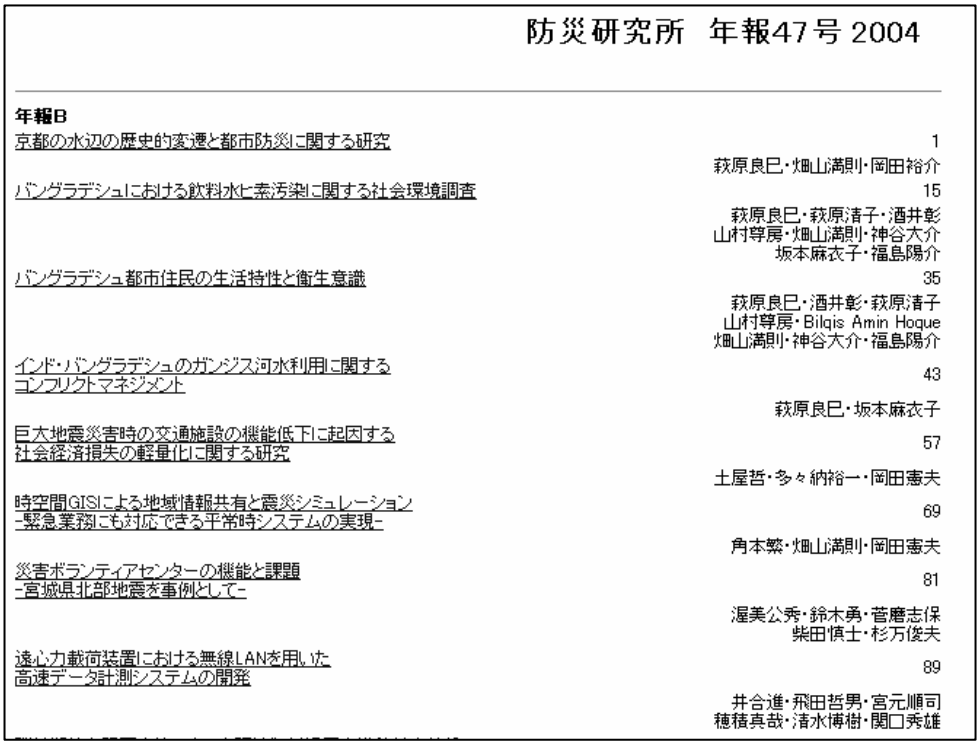


図５　文献閲覧用目次 HTML

４．　防災検索システム改良案

４．１　目次検索

今年度で、年報第１号～第４７号まで、既刊されているほとんどすべての年報文献を電子ファイル化、掲示可能となった。さらにブリテン第２０号～第４５号、公開講座第１回～第１４回分も追加可能である。ここで新たなる問題が生じる。現在の目次検索は、図５のように毎年ごとの冊子別目次へのリンクが並んでいる。青い「A」、「B」、「C」といった冊子のボタンをクリックすることによって、図６のような「目次からの年報文献検索ページ」へリンクする。前述の新たなる問題とは、号数や文献が多数になるにつれて、ユーザが閲覧したい文献に辿り着くまでの時間が増大することである。

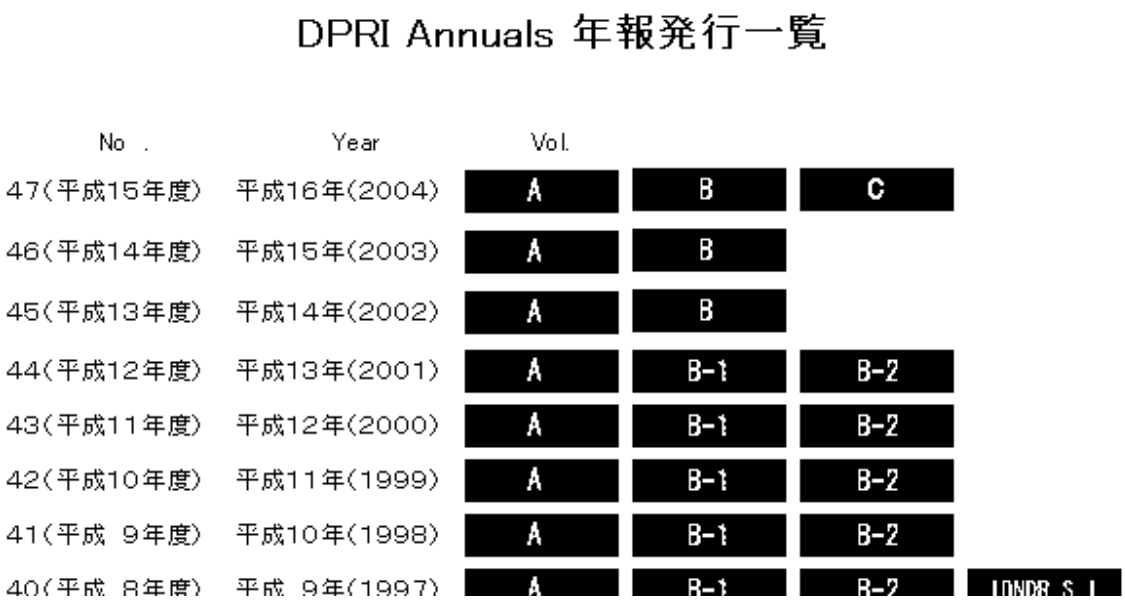


図６　目次一覧

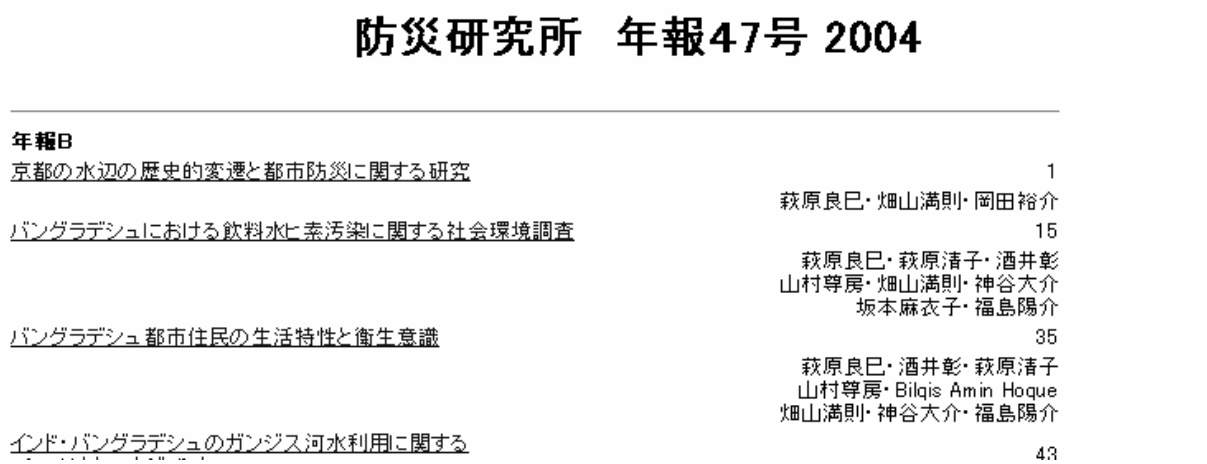


図７　目次からの年報文献検索

この問題の解決法として、現在二つのことが考えられる。

１つは、JavaScript、Flash を使用して、新たなるメニューを作成することである。図８は、JavaScript を利用したメニュー式の目次検索で、「文献の種類」→「出版年」→「冊子の種類」の順に選択することができる。これにより、現状の目次システムよりユーザの負担を減少させることができると考えられる。

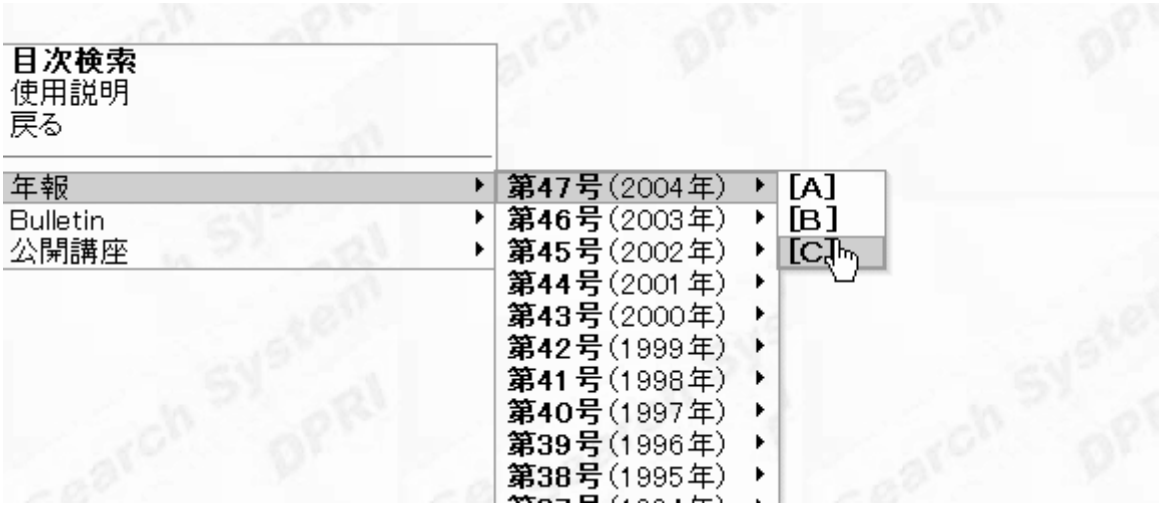


図８　JavaScript を利用した目次検索

４．２　キーワード検索

現在、キーワード検索システム（カテゴリ検索システム、簡易検索システム）は、主に「タイトル」、「著者名」、「出版年」、「キーワード」等から「年報」、「ブリテン」のみを対象に絞込み検索できるシステムになっている。しかし、防災研究所が蓄積してきた防災研究資料は、定型のドキュメントだけではなく、ハザードマップ、災害写真等の静止画像、災害映像、音声、地震波形、地理情報等、種類は多岐に渡る。そのため、現状のシステムでは、定型の防災資料ドキュメントしか検索することができないシステムであり、今後はさらに改良したシステムが必要であると考えられる。そこで、防災研究資料に共通して存在すると考えられる、「タイトル」、「著作権」、「概要」の三つを柱としたシステムを検討している。

４．３　全文検索システム

現在の全文検索システムは、オープンソフトである日本語全文検索「namazu」を活用したシステムとなっている。検索システムプログラム自体はすでに完成されており、変えることはできないが、検索結果情報の充実は可能である。そこで、従来の検索結果情報に、「著作権」、数行までの「概要」を表記することを検討している。改良されれば、検索結果の中から、ユーザが求める防災情報をより効率よく提供できると考えられる。

５．　プロジェクトの最終結果報告と今後の課題

既刊された年報すべて、ブリテン第４５号～第２０号、公開された公開講座すべての資料の電子ファイル化、及び、これら防災研究資料を Web 上で高速に検索できるシステムの構築が終了致した。

今後は、膨大なデータのバックアップと管理体制の充実、４で述べた改良案を元に、より良い防災情報検索システムの作成と XMDB、自己点検等の他の検索システムとの連携を目指していく。

６．　終わりに

限られた予算の中、年報だけに留まらず、ブリテン、公開講座まで電子ファイル化及び防災資料検索システム実用化までに至りました。

３年という長くも短い間、続けてきた COE プロジェクトでしたが、ようやく１つの区切りがつき、これからの防災資料の電子ファイル化の非常に大きな第一歩をようやく踏み出せた気が致します。

ひとえに室長をはじめ、支援頂いた技術室の皆様、教員の方々、そして本プロジェクトの影の主役であり、一番の功労者である電子ファイル化作業を手伝って頂いた非常勤職員の皆様の長く地道な努力の賜物と感じております。ここに記して心より厚く御礼申し上げます。