

データベース構築促進及び技術開発に関する報告書

新聞縮刷版見出しデータベースの構築

平成 5 年 3 月

財団法人 データベース振興センター
委 託 先 株式会社 朝 日 新 聞 社

本事業は、日本自転車振興会から競輪収益の一部である機械工業振興資金の補助を受けて作成したものである。

データベースは、わが国の情報化の進展上、重要な役割を果たすものと期待されている。今後、データベースの普及により、わが国において健全な高度情報化社会の形成が期待される。さらに海外に対して提供可能なデータベースの整備は、国際的な情報化への貢献および自由な情報流通の確保の観点からも必要である。しかしながら、現在わが国で流通しているデータベースの中でわが国独自のものは1/3にすぎないのが現状であり、わが国データベースサービスひいてはバランスある情報産業の健全な発展を図るためには、わが国独自のデータベースの構築およびデータベース関連技術の研究開発を強力に促進し、データベースの拡充を図る必要がある。

このような要請に応えるため、(財)データベース振興センターでは日本自転車振興会から機械工業振興資金の交付を受けて、データベースの構築および技術開発について民間企業、団体等に対して委託事業を実施している。委託事業の内容は、社会的、経済的、国際的に重要で、また地域および産業の発展の促進に寄与すると考えられているデータベースの構築とデータベース作成の効率化、流通の促進、利用の円滑化・容易化などに関係したソフトウェア技術・ハードウェア技術である。

本事業の推進に当って、当財団に学識経験者の方々に構成されるデータベース構築・技術開発促進委員会(委員長 山梨学院大学教授 蓼沼良一氏)を設置している。

この「新聞縮刷版見出しデータベースの構築」は平成4年度のデータベースの構築促進および技術開発促進事業として、当財団が(株)朝日新聞社に対して委託実施した課題の一つである。この成果が、データベースに興味をお持ちの方々や諸分野の皆様方のお役に立てば幸いである。

なお、平成4年度データベースの構築促進および技術開発促進事業で実施した課題は次表のとおりである。

平成5年3月

平成4年度 データベース構築・技術開発促進委託課題一覧

分野	課題名	委託先
社会	1 変異タンパク質配列データベースの構築	日本電子計算(株)
	2 新聞縮刷版見出しデータベースの構築	(株)朝日新聞社
	3 ファジィに関する文献データベースの構築	(財) 日本情報処理開発協会
	4 医療用医薬品抗生物質データベースの構築	(株)小田島
	5 交通事故調査データベースの構築	(財) 日本自動車研究所
	6 楽器データベースの構築	(株)ダイソメディアサービス
	7 人体計測データベースの構築	(社) 人間生活工学研究センター
	8 大学におけるデータベース利用教育システムのプロトタイプ作成	日外アソシエーツ(株)
	9 先進複合材料データベースの構築	(財) 次世代金属・複合材料研究開発協会
	10 博物館所蔵地図資料所在情報データベースの構築調査	(財) 地図情報センター
中小企業振興 地域活性化	11 地域流通最適化データベースのプロトタイプ作成	(社) 日本ボランティア・チェーン協会
	12 異分野研究のための知的オリエンテーション・データベースシステムのプロトタイプ作成	(株)けいはんな
	13 在宅勤務者サポート・データベースの構築調査	(株)志木サテライトオフィス・ビジネスセンター
海外	14 銅基複合材料日本特許英文データベースの構築	神鋼リサーチ(株)
	15 技術協力供与機材データベースのプロトタイプ作成	(財) 日本国際協力システム
	16 先端産業分野における専門用語の電子辞書データベース化の調査研究	科学技術情報研究所(株)
	17 マーケティングコードの英文データベースの構築	(株)帝国データバンク
技術	18 安全研究における多重ソース・システム構築のための基本安全用語データベースの開発	(株)紀伊國屋書店
	19 3次元マッピングデータベースの技術開発	(株)日本総合技術研究所
	20 データベース検索サポートシステムの調査研究	セントラル開発(株)情報図書館 RUKIT
	21 グループウェアにおけるデータベースシステムに関する調査研究	(株)イフ・アドバタイジング
	22 パーソナルコンピュータとLANの利用による非定形データベースのプロトタイプ作成	(株)メイテック
	23 知的資源型データベースの調査研究	(株)ジャパンコミュニケーションズ ズインスティテュート

目 次

1. はじめに	1
1.1 見出しデータベースの目的	1
1.2 新聞記事データベースとは	3
2. 仕様の検討	10
2.1 データの基本構成	10
2.2 データの取捨選択	13
2.3 データ構造の決定	14
3. データづくり	20
3.1 データ起こし	20
3.2 光学読み取り機の導入	20
3.3 校正作業	29
4. システム構築	31
4.1 簡易システム構築	31
4.2 全文フルサーチ型検索システム	32
4.3 H I A S K編集システムの改造	35
5. 検証	38
6. 今後の課題	48

1. はじめに

1.1 見出しデータベースの目的

パソコン通信の利用者数が1992年には延べ100万人を超え、オンラインによる各種データベースの利用はごく身近なものになってきた。

朝日新聞社では1984年8月からニューメディア本部の担当で新聞記事データベースの構築を開始し、1986年4月から平和情報センター（現在のジー・サーチ）と日本アイ・ビー・エムを皮切りに記事データベースの一般提供を始めた。

1993年2月現在、ジー・サーチ、日本電気のC&C-VAN、日経テレコン計3社のディストリビューターにデータを提供、これにゲートウェイでアスキーネット、PC-VAN、NIFTY-Serve、アサヒパソコンネットなどのネットが接続しており、これらのIDを合計すれば百万にもなろうとするユーザーが事務所や家庭から電話線を通じてオンラインで新聞記事データベースを利用できる状況にある。

しかし、当社が記事データベース収録を開始したのは1984年からであり、さらに諸々の事情から一般提供しているデータは1985年1月以降のものである。また、収録しているデータも当初は東京本社発行紙面のうちのニュース面だけであった。

その後、家庭面、文化面、読書面などへ順次収録範囲を拡大してきたが、スポーツ面は93年1月から収録を開始するなどデータのやや不完全なものがあつた。

さらに、1984年7月以前の記事を探すには、社内利用においても、調査部の切り抜き、あるいは各月に発行される縮刷版を繰って見る等しか検索の方法はなかった。

ユーザーが記事データベースの利用に慣れてくるにつれ、昔のデータを何とか同じような方法で検索できるようにならないかという要望が内外から寄せられるようになった。

しかし、データベースを過去にさかのぼって起こすにはたやすいことではない。いちいち手入力するのでは膨大な人手と手間がかかる。全文型データベースを起こすにしても、過去の記事にそんなに需要があるものだろうか。そこで着目したのが当社が毎月発行している縮刷版であった。

朝日新聞記事縮刷版は、大正8年（1919年）創刊の歴史をもち、爾来1年も欠かすことなく、毎月発行されてきている。しかも、巻頭には記事の見出しをそのま

ま項目ごとに整理した詳細な記事索引がついている。

ところで、実際に縮刷版で過去の記事を探そうとしたら、それがなかなか大変な作業であることに気がつくだろう。

目的の記事がいつ掲載されたかわかっている場合はいい。しかし、正確に事件の起こった日付を覚えているケースはむしろまれであろう。そうすると、各号の巻頭にある索引を利用することになる。

しかし、索引は1カ月単位でしか探せない。1000ページを超し、数キロもある縮刷版を次から次へと探すのはかなりの肉体労働である。開架式の図書館ならともかく、閉架式の図書館だとその困難さはさらに増すことだろう。

この記事見出し索引をデータベース化し、見出し文のほか、発行日、朝夕刊の区別、面名、縮刷版における掲載ページなど、記事の所在情報を示すものをつくれば、資料探しの手間は飛躍的に改善される。長期間を通覧できるものができたら、年表、事件史といったこれまでにない新しい用途も開けるのではないか、こういった構想で構築に着手したのが縮刷版見出しデータベースである。

1.2 新聞記事データベースとは

ここです、新聞記事データベースとはどんなものであるかを概観しておきたい。

朝日新聞記事データベース（H I A S K＝ハイアスク）は、前述したように1984年から記事の収録を開始した。

記事本文中の名詞を自動インデクシングでキーワード（KW）として切り出し、シソーラス（統制語）は不要、記事文中の思いついた言葉で検索してください、というひきやすさ、そしてヒット率の向上をめざした。

朝日より先に新聞記事データベース構築に着手していた社に日本経済新聞社と中日新聞社があったが、手動によるKW付け作業であり、自動記事インデクシング、自由語検索を採用したのは朝日が最初である。フルテキスト（抄録でなく全文収容）とKW自動抽出、自由語検索というこの方式は幸い各社の追随するところとなり、日本の新聞記事データベースの主流となっている。

H I A S Kで採用している自動インデクシングシステムはA J A X（エイジャックス）といい、日本アイ・ビー・エムと共同開発したものである。これは後に、日本アイ・ビー・エムが市販している日本語情報検索システムJ A I R Sのインデクシングシステムとして製品化された。

ホスト機としてI B Mの3081-GXを専用機とし、データベース編集端末には同じくI B M5550型26台をT S O直接端末として使っている。処理する記事件数は1日平均約500本（ほかに大阪本社でも150本を処理）。84年からの記事総蓄積本数は約97万5000件（93年2月末現在）にのぼる。

記事データベースの素材、元データは最初からできているといってよい。これは、元データから起こさねばならない縮刷版データベースとのとの違いである。いまではどこの新聞社でも当たり前のことになってしまったが、朝日新聞はコンピュータによる紙面制作（C T S＝Computerized Typesetting System）を採用している。この中に必要な文字列情報は含まれている。

しかし、新聞の組版情報には記事型データベースには不要な情報も多く含まれている。例えば写真、カットなどのイメージデータ、文字の大きさ、書体、文字の位置などの情報である。また、記事本文も一つのテーマが一つのファイルにまとまってくればよいが必ずしもそううまくはいかない。

そこで、夜のうちに前処理を行う。通常この作業を内部用語で「記事ばらし」と呼んでいる。

まず、記事をテーマごとにまとめる。新聞制作では見出し、前文、記事本文などはそれぞれ独立したIDで運用されており、グルーピングという編集支援情報をもとにこれを1本の記事に結合する。

次に発行日、ページ、面名、発行本社（東京本社以外、大阪、西部、名古屋各本社発行の記事も収録している）、出稿部、著作権の有無、写真の有無などの書誌的事項を抽出する。これは主に紙面制作管理情報から自動発生させる。

それから「漢字カナ変換」といって、一種の文法解析を行う。

「漢字カナ変換」では、まず切り出し語辞書をもとに、句読点、ハイフン、感嘆符、疑問符、符号、定型句などで文を大きく分割する。

次に単語・付属語辞書と照合し、さらに漢字とカタカナ、ローマ数字、アラビア数字、ローマ字など異なる字種の前で単語を切る。さらに漢字カナ辞書と照合し、読み・文法情報をつける。ここまでは夜のうちに自動化されたバッチ作業で行われる。

次に「編集作業」を行う。

「編集作業」は約26人のスタッフで日曜祝日を除く連日作業している。

「編集作業」の主なものはまず書誌的事項のチェック。「写真の有無」（図版、カット、写真の区別が自動的にはできない）、「著作権の有無」など。人間の判断が入らないとできないところを重点的に行う。

次に見出しを手直しする。「カット」は白か黒かの画像情報だけで、文字列情報を含んでいないため、見出しから欠落してしまうからである。さらに短信のように編集上の制約から字数を制限しているものは主語・述語がはっきりするように直す。事件の続報とか受け記事で必要なKWが文中に欠落している場合は意図的にこれを補う。

次に記事に分類語をふる。これは朝日新聞が独自に考案したもので、記事の性格によって大、中、小あわせて約820項目の分類語のどれかに記事を割り当ててしまうものである（93年からはこれを約100項目に簡素化した）。一種のシソーラスに近い。

H I A S Kでは文中の名詞をKWにする自由語検索と分類語検索の2本立てで検索できるようになっている。理由は雑音の除去と精度の向上のためである。

<朝日新聞記事データベース 仕事の流れ>

●ネルソン

データベース作成 IBM3081-GX

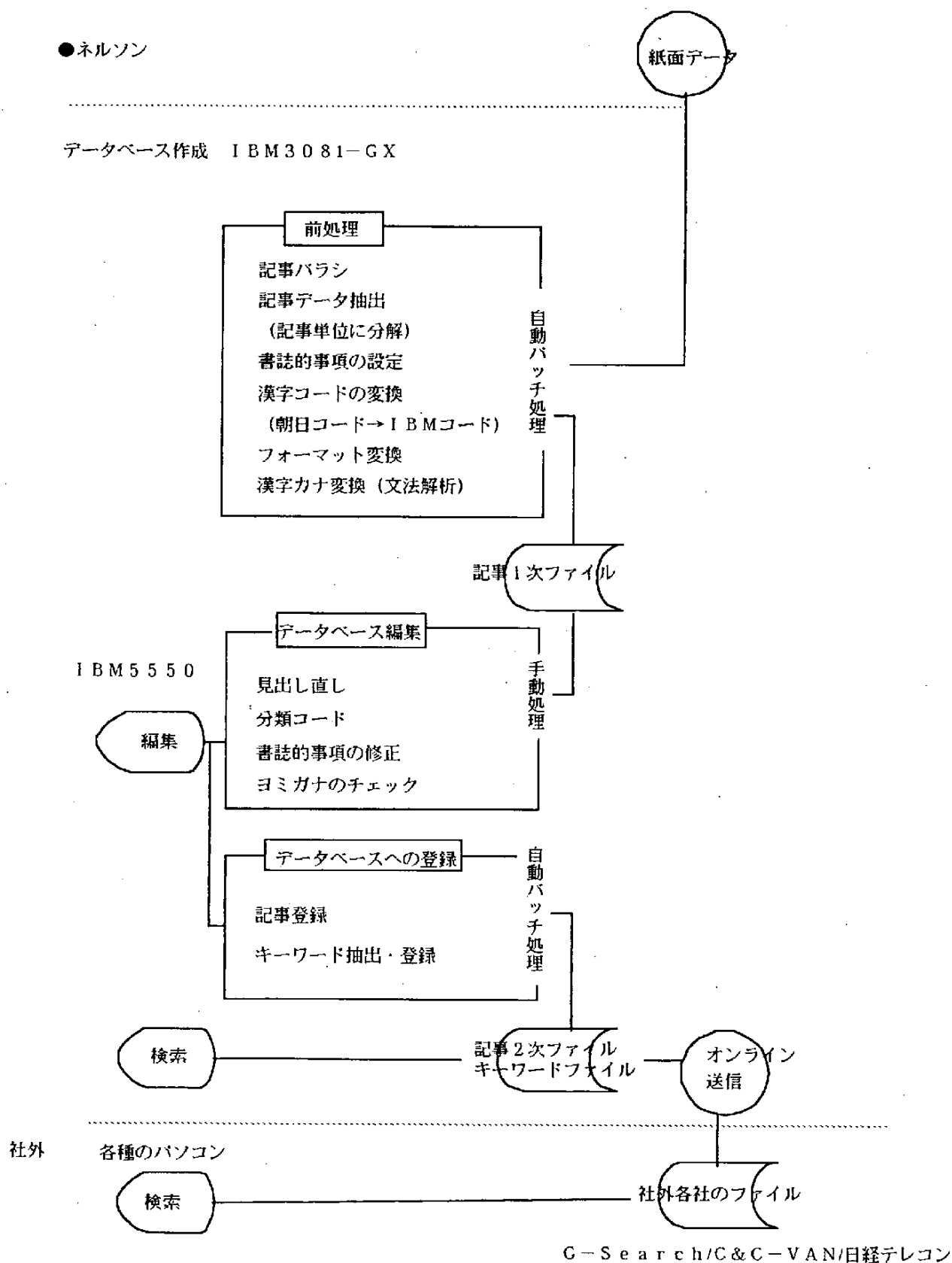


図1-1 新聞記事データベース編集の流れ

例えば「火事」の記事を探じたいとき、利用者は「火事」「火災」をKWにするとと思われるが、実際の新聞記事には「出火」「全焼」「焼死」などの言葉は出るが、「火事」という言葉はあまり使われないものである。H I A S Kは記事文中からKWを抽出しているから、これらの言葉が使われていなければ、検索対象からは漏れてしまう。また「カジ」では「家事」の記事もヒットしてしまう。こんなとき、分類語として「火災」をつけておけば、こういったわずらわしさをなしに「火事」に到達できるというわけである。

次に読みのチェックを行う。

先に述べた「漢字カナ変換」で文法解析のあらかたは終わっているわけだが、固有名詞とくに初出のものについては、辞書になく、従って読み損ねをしていることがままある。カタカナKWで検索する利用者のためにはこれは致命的なので怪しそうなものにはチェックをいれる。

これはディスプレイ画面で該当する単語のところにカーソルをもっていき、特定のキーを押すとその読み方が表示される仕組みになっている。

読み方が間違っているとまず、正しい読み（カタカナ）を入れ、それにあわせて漢字を入力する。文法情報はこのとき自動的に修正させる。

読み方のわからないものは出稿部に問い合わせる、場合によっては直接取材して確認する。

これと同時に不要なKWの選択を行う。明らかに不要と思われるKW部分は切り出し対象領域から外してしまうのである。

こうした作業を経て、編集が終わるとプリントアウトしてデスクチェックの後、KW切り出しにかける。

KW切り出しは紙面1個面を単位にして行う。対象記事IDをリスト表示画面に呼び出し、コマンドをうつと、文法情報に従って指定対象領域から名詞の品詞情報がついた文節、カタカナ文字列、ローマ字文字列、単独でなく前後に単語が付随している数字などをKWとして切り出す。ここで漢字KWとカタカナKWの両方が同時に形成される。

KW切り出しが終わったデータは一時蓄積ファイルに転送する。この時複合語の展開を行う。そして各ディストリビューターにオンライン送信する一方、内部では毎晩データベースを更新する。

人手を加えないでKW切り出しをおこなった場合の精度は96%前後。編集に人

手を介在させるのはこの96%をいかに100%に近づけるかという作業であるといっても過言ではない。

こうした新聞記事データベースと同じ編集方式で、見出し総覧ともいえるべき縮刷版見出しデータベースを構築すれば、検索方法も共通で利用者の利便も図れ、これまで効率的利用が困難だった電子化以前の記事の有効利用に役立つと考えたわけである。

ただ、H I A S Kのシステムは新聞記事をデータベースに取り込むことを主眼としたシステムで、ワープロ文書やMS-DOS形式の文書がストレートに取り込めなかったため、中規模のシステム改造を行った。

. 文書名 T930125M20--08
 . 発行日 930125
 . 紙誌名 N
 . 朝夕刊 M
 . ページ 20
 . 版名 14
 . 面名 SP2
 . 発行社 T
 . 出稿社 T
 . 出稿部 ウンドウ
 . 写真 Y
 . 文字数
 . 著作権
 . 位置 01D35L

見出し

最年少大関貴花田“バブル昇進” 審判部、苦しい言い訳

分類

. CM @--、武道
 . CM @K-、データ
 . CM @C-、日本

本文

千秋楽で完敗した貴花田が大関に昇進する。二十歳五カ月の史上最年少大関。昇進前三場所の計35勝は、確かに数字的には大関として問題ない。しかし、上位陣が手薄なうえ、平幕に3敗、そして曙に全く相手にされなかった勝負を受けての昇進は、人気をあてこんだ協会側の興行優先政策以外の何物でもない。これではあまりに本人も惨めだ。

とくに昇進を決定づける審判部の“言い訳”が苦しい。これまでは昇進の条件としては、前三場所の勝ち星はあくまで参考で、相撲内容を重視するとの姿勢を強調。この日の取組前に行われた三賞選考委員会の席上でも鏡山（元横綱柏戸）、陣幕（元横綱北の富士）両審判部長は「結びの一番の相撲内容を見て、昇進を考える」といった。

KW文書サンプル

020 . *見出し*	
020 最年少・大関・貴花田	サイネショウ・オセキ・タカハタ
020 バブル・昇進	ハブル・ショウシン
020 審判・部	シンパ・ソ・ブ
020 言い訳	イワケ
030 . *分類*	
030 @--、武道	@--.87
030 @K-、データ	@K-.08
030 @C-、日本	@C-.2104
040 . *本文*	
040 千秋楽	センショウラク
040 完敗	カンバイ
040 貴花田	タカハタ
040 大関	オセキ
040 昇進	ショウシン
040 二十・歳・五カ月	ニジユウ・サイ・ゴ・カゲツ
040 史上・最年少・大関	シジョウ・サイネショウ・オセキ
040 昇進・前・三・場所	ショウシン・ゼン・サン・バシヨ
040 計・35・勝	ケイ・35・ショウ
040 数字・的	スウジ・テキ

図1-2 データベースの検索文書とKW切り出し例

040 大関	オオセキ
040 問題	モンダイ
040 上位・陣	ジョウイ・ジン
040 手薄	テウス
040 平幕	ヒラマク
040 3・敗	3・ハ・イ
040 曙	アケボノ
040 相手	アイテ
040 勝負	ショウフ
040 昇進	ショウシン
040 人気	ニンキ
040 協会・側	キョウカイ・ガリ
040 興行・優先・政策・以外	コウギョウ・ユウセン・セイサク・イガイ
040 何物	ナニモノ
040 本人	ホンニン
040 惨め	ミジメ
040 前・三・場所	ゼン・サン・ハ・ジョ
040 勝ち星	カチホシ
040 参考	サンコウ
040 相撲・内容	スモウ・ナイヨウ
040 姿勢	ジセイ
040 三賞・選考・委員会	サンショウ・センコウ・イインカイ
040 席上	セキジョウ
040 鏡山	カガミヤマ
040 元・横綱・柏戸	モト・ヨコヅナ・カシワバ
040 陣幕	ジンマク

2. 仕様の検討

2.1 データの基本構成

朝日新聞縮刷版見出しは図2-1のような仕組みになっている。

まず記事の種類によって「国土・人口」、「政治」「経済」などカットつきの大項目の見出しがついている。その下に「開発・保全」「気象」、政治関係でいえば「政党」「国会」などの中項目的見出し、さらにその下に「土地対策」「環境保全」といった小項目的見出し、さらにその下に「国土利用計画法」といった具体的各論の見出し、その下に各見出しが並ぶという構造になっている。各見出しの下に和数字がその号に掲載されているページ数を示し、洋数字が何段目にあるかを示す。

縮刷版見出しデータベース構築に当たってまず直面した問題は、限られたデータの中からいかに過不足なく有効な情報をひき出せるようにするかであった。

縮刷版見出しは全紙面の見出しをそのまま拾い上げ、項目別に整理したものである。新聞の見出しは記事の中身を凝縮して伝えるものだが、表現上字数の制約をもっている。

一般的には3段以上の大きな見出しの場合、太い見出しは7-9字、わきの細い見出しは9-10字の組み合わせが読みやすいとされている。

科学的に検証したわけではないが、大手紙の紙面をみても不思議に似通っている。人間が新聞を一瞥したとき、一瞬に内容を把握できる字数には限度がある。30字、40字を使い内容を詳細に説明するよりも、結局、先に述べたような字数がちょうどよいのだという説明を整理部の担当者から聞いたことがある。

また、短信記事では極端に見出しの字数が制限されていて、主語すら省略されているものがある。

このように見出しは極端に切り詰めた表現をしているので、検索キーワードが非常に限られてしまう。このことは、検索にあたっては必要な情報が漏れる、逆に似たような情報が多数ひっかかってゴミだらけという、両方の欠陥を招く恐れがあった。

例えば、「宮沢首相の改憲否定発言」を探すとする。見出しではまず、首相の名前は省略してしまうから、たいがいは「首相、改憲を否定」といった類の表現になるだろう。同じ見出しスペースなら、字数が少ないほど大きな活字が使える、

索引目次

週間報告・座標	3
国土・人口	3
国土・資源・人口	
<開発・保全>	3
国土開発・都市計画	
土地対策・環境保全	
<気象>	3
天気・台風・地震・	
火山・天文	
政治	4
憲法・人権問題・政	
治思想団体・平和運	
動	
<政党>	4
政治資金・政党別動	
向	
<国会>	5
衆議院・参議院	
<法令・法案>	5
<選挙>	5
選挙制度・国会議員	
選挙・地方選挙・選	
挙違反	
<内閣>	6
首相・閣僚・閣議・	
各省庁・人事・白書	
<財政>	7
予算・税・財政投融	
資・地方財政	
<国防>	7
日米安保・防衛計画	
・自衛隊・在日米軍	
外交	7
領土領海・条約協定	
・対外援助・各国別	

<国交>	9
承認・断交・大公使	
・要人の来日・出入	
国査証・在日外国人	
・国際親善	
経済	9
経済政策・景気動向	
・国際収支・対外経	
済協力・公取委・財	
界・経済団体・会社	
企業・決算・倒産	
<金融>	10
金融政策・金利・銀	
行・外国為替・保険	
・証券	
<産業>	11
産業政策・各産業	
(エネルギー・鉱工	
業・農林水産)	
<流通・商業>	13
卸小売業・百貨店ス	
ーパー・総合商社・	
倉庫業	
<物価>	13
物価対策・価格動向	
・物価指数・料金	
<貿易>	13
輸出入動向・輸出・	
輸入・見本市・関税	
運輸・通信	14
交通・道路・鉄道・	
航空・海運・通信・	
郵便	
労働	14
春闘・雇用・失業・賃	
金・労働組合・争議	

世界	15
<国連>	15
国際会議・国交・国	
連・軍備軍縮・国際	
経済・金融・石油	
<アジア>	17
国別(各州とも同じ)	
<中東>	20
<アフリカ>	20
<ヨーロッパ>	21
<アメリカ>	22
<オセアニア>	24
文化	24
文化交流・学術・著	
作権・国語・文化財	
記念物	
<科学>	24
みんなの科学・くら	
しの科学・計測・発	
明発見・観測調査・原	
子力・動物植物・医学	
<教育>	25
教科書・学校・教職	
員・学生・宗教	
<学芸>	26
寄稿・文学・美術・	
音楽・映画・芸能・	
演劇・舞踊・能・狂言	
<出版・報道>	28
新聞・出版・図書特	
集・放送・広告	
社会	28
皇室関係	
<一般>	28
風俗・善悪善行・祝	
日・年中行事・戦争	
関係・宝くじ	
<人室>	29
叙位叙勲・結婚離婚	
・死去	

<公害>	30
公害対策・地盤沈下	
・大気汚染・水質汚	
染・農薬汚染・公害	
病	
<事件・事故>	30
中絶・自殺・爆発・	
遺失・交通事故・航	
空事故・火事・災害	
<犯罪>	32
防犯・過激派事件・	
誘拐・密輸・詐欺	
横領・汚職・偽造模	
造・傷害・暴力おど	
し・殺人・強盗・窃	
盗	
<司法・警察>	33
法曹界・裁判所・行	
刑・検察・警察	
<裁判>	33
訴訟・公判・判決	
国民生活	34
家計・消費者・消費	
者運動	
<衣・食・住>	34
<保健・衛生>	34
病気・医療・医薬品	
・衛生	
<福祉・厚生>	34
社会保障(健保・年金)	
・老人問題・身障者	
<婦人・青少年>	35
<観光・娯楽>	35
レジャー・余暇・施設	
・趣味・娯楽・競馬	
家庭面	35
ひととき・服飾・美	
容・料理・住まい・	
園芸	

凡例	
◎各見出し語の下に和数字	
はページ数、洋数字は見出	
し語のある段数を示す。	
◎見出し語の分類は一語一	
か所方式をとっているの	
で、引いた項目にない場合	
は、関連のある他の項目を	
参照されたい。	
◎矢印(⇒)は参照項目を	
示す。	
◎2日付朝刊は休み	
特集面	36
解説面・月曜ルポ・	
にゅうすらうんじ	
<日曜版>	37
日本史の舞台・ぶっ	
くす・紙上創作展・	
暮らしと健康・遊び	
スポーツ	38
<外国>	38
種目別	
<国内>	38
種目別	
地方	41
地方自治・首都圏	
<東京都>	41
一般・都政・都議会	
・23特別区・市部・	
郡部・島部	
写真	42
読者の新聞写真コン	
クール・写真特集	
本社	42
社説・今日の問題・	
天声人語・声・論壇・	
世論調査・社告・事業	

週間報告	12月28日・1月3日・一七8〇
1月4日・10日・三六五8〇1月	
11日・17日・六二五8〇1月18日	
24日・八九五8	
座標	
生活の豊かさ・都留重人・三二九	
1〇頁さんの家庭論・宮下展夫・	
六九七1〇過熱する生命技術・柴	
田鉄治・八五九1	
国民の意識(本社世論調査)	
生活苦高まる声・自衛隊で	
改悪賛否ほぼ同率・六五七	
本社調査に見る国民意識	
不満多いが安定を評価・七四1	
暮らしと政治・憲法と自	
衛隊・老後の暮らし・エ	
ネルギー源・七四1	
仲よくする国・七四9	
けなげな女性が理想像・好	
きな女性性・女性の進出	
・生まれかわるなら・七五1	
どちらの名字・どこまで	
信用・七五5	
日本人らしい平衡感覚	
建前と本音をたぐみに区	
別・七五6	
調査方法・七五9	
国土・人口	
落ち込む出生率・昨年は史	
上最低・両午の41年下回	
る・三1	
全国で86万人・年男・年女・三9	
人口動態	
社会増員立つ中小都市・人	
口動態・移動率は7%割	
る・一八七8	
世帯構成(55年厚生行政基礎調査)	
家庭の老齢化・ミニ化加速	
高齢者世帯・24万超す	
この10年で倍増・五九1	
(表) 60歳以上の者のい	
る世帯の構成割合の推	
移・五九2	
(表) 人員別にみた世帯	
数と平均世帯人員の動	
き・五九4	
開発・保全	
【国土開発】	
定住構想	
高齢者を念頭に整備・「地	
域づくり」で定住構が報	
告書・八九二6	
【都市計画】	
都市計画道路変更案を了承	
都計審・一一〇一8	
【土地対策】	
首都圏の土地利用状況・10	
以平方メートル航空写真に	
建設省が新年度着手・六七八1	
宅地化へ基礎資料・六七八1	
【環境保全】	
環境アセスメント法は「法令・	
法案」	
清流復活北歌へ紹介・環境	
庁・区民の知恵新賛・江	
戸川区の親水公園・四六一1	
環境保全「ナショナル・セ	
ラスト」導入・七三三1	
土地買取りの住民に税	
金優遇・環境庁方針・七三三1	
あくまで住民主体・七三三6	
湖沼保全で答申	
水辺に憩いの場を・湖沼保	
全・中公審が答申・市街	
化に歯止め・九六七1	
【解説】汚濁・深刻な状況	
省庁の課題法整備を九六九1	
湖沼保全の答申内容・要	
旨・九六九4	
気象	
【災害】	
大寒気団ひかえ81年始動一三七5	
北陸地方にまた寒波・一七〇6	
今冬は「第3寒波」・気象庁	
が警戒訴え・二〇九8	
なお大寒の恐れ・寒気団が	
また接近・二二19	
大寒・平野部は一服・内陸	

図2-1 縮刷版記事索引のサンプル

その方が読者に与えるインパクトが強いからである。

ところが「改憲否定発言」をした首相は歴代何人もいる。むしろ、歴代首相のすべてが建て前としては「改憲」を否定してきた。そうだとすると「首相」「改憲」「否定」というキーワードでこのテーマを検索したら、歴代すべての首相発言がヒットしてきかねない。

逆に、あるテーマを広範に探したいときの漏れも予測された。ここ10年間の地震の年表を作成したいとする。ところが新聞の見出しは「石垣島で震度1」、釧路沖地震のように大規模なものは「釧路で烈震、M7. 8」というように「地震」というキーワードでは一括してくれないものが結構多い。

新聞独特の受け記事の見出しも悩みの種であった。例えば内閣が交代した時の野党の談話「新味無し ○○委員長語る」などである。これはこの記事の前の方に内閣改造の本筋を報道する記事があり、それを前提にしての表記であるからこんな表現も許されるのだが、一本に独立したデータとして扱う場合は欠陥だけである。

予想される問題点を洗い出してみた結果、「大」「中」「小」とついている各項目見出しが、H I A S Kの概要の所で述べた分類語として使えないか検討してみた。縮刷版の目次をみてもわかるように、各項目見出しの構造は上下関係が比較的しっかりしている。「政治」とやれば、その項目の下の全見出しが集まってくる仕組みである。

しかし、検討を進めるにつれ、この大項目や中小項目の見出しは時代によってかなり変動があることがわかった。時代とともに一つのテーマが項目間を移動しており、一貫したものになっていなかったのである。これでは分類語での検索に漏れが出る。

結局、各論見出しや大中小各項目の見出しをを上下関係無しに並立するキーワードとして加えてやれば、それなりの絞り込みの役には立つのではないかという結論に達した。検索の精度向上には役立とう。先ほどの野党党首談話でも、項目見出しに「××内閣成立」という項目があれば、特定がつく。

2.2 データの取捨選択

縮刷版のデータはすべて取り込むことを基本方針とした。しかし、心ならずも省略したものもいくつかある。

商況、天気予報の類である。これは縮刷版の表記では

株価

東京株式第1部・第2部＝一一一1－九九七1

外国株＝一一6－九九七5

天気

今日の天気＝一10－九六－10

というように月初めと月末の掲載ページしか表記されていない。

内容を示す手がかりもなく、東京株式市況の項目が連続して並ぶのは積極的意味がないと判断して、このような形式のものはほとんどを削除した。

ただし、「天声人語」といった、定型コラムは同じような形式であったが、一応収容した。

2.3 データ構造の決定

データ構造と表示の形式は図2-2-1のようにした。

見出し表示は2行で構成され、1行目は見出し本文、2行目が発行日、朝夕刊の区別、掲載面、そして縮刷版の掲載ページ、掲載段という所在情報を示す。

現行の朝日新聞記事データベースとの整合性を図るため、見出し表示は1行36字以内とした。そのため、長行の見出しによってはおさまりきれないものが出てくるので、それは本文表示（詳細表示）という形式で救済をはかることにした。

逆に訃報のように「何野何兵衛氏」としか表示がないものは、その上位項目の見出しをつけ加えて「何野何兵衛氏_訃報」という注釈めいた表現をとれるようにした。

本文表示（詳細表示）は、2行では表示しきれない部分の情報を示すものだが、36字を超えた見出し部分の残りと、さらに、その上位にどんな項目が含まれているかあわせて表示するようにした。

項目1はカット付きの大項目、項目2は波形罫線つきの中項目・・・と階層をそのまま表わす方式とした。

データはMS-DOS形式のファイルとして起こした。ワープロでは機種によってフォーマットが違うので、MS-DOSが一番標準的であろうと判断したためだ。起こした文書は、図2-2-2のような書式になっている。

カット付きの大項目は | |、波形罫線つきの中項目は< >、【 】で囲まれた小項目は[]、ゴチックの見出しは頭付き、各見出しの本文は1字下げという構造である。

これをトリガーにしてQuick Basicを使ったパソコンの変換プログラムにかけ、各上位項目を見出し文に自動的に付加させた。

見出し文の前にゴチック見出しをつけ加え、さらに小項目、中項目・・・と足していき、見出し1本がそれぞれ改行で終わる1レコードとして構成される仕組みにした。

次に掲載日、掲載面の設定である。

元データの見出し文の最後にある和数字はその縮刷版における掲載ページを表している。掲載年と掲載月はその縮刷版で特定されるから、これを手がかりに、掲載日、掲載面を自動発生させることにした。

まず、図2-2-3のような変換テーブルを作成する。1日の朝刊は○ページから○

縮刷版 D B 表示形式

【1レコードのフォーマット】

半角スペースがデータの区切り、空白データは"<>"、改行がレコードの区切り。

8412-0687	841230	M	2	1096	5	政治	財政	予算	60年度予算政府案	反響	努力は評価
↑	↑	↑↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
ID	年	朝面	縮	段	項	項	項	項	項	項	見
(年月+通しNO)	月	夕刊	刷	目	目	目	目	目	目	目	出
	日	紙	の	1	2	3	4	5	5	し	
		面	頁								
		頁									

8412-0687	年月とその月の縮刷中の通し番号、IDとして使用、9バイト。
841230	その記事の掲載年月日、6バイト。
M	朝夕刊、M=朝刊、E=夕刊、1バイト
2	面、紙面中での頁、1～2バイト。
1096	縮刷版での頁、1～4バイト。
5	段、紙面中の位置、"1"-"15"の1～2バイト。
政治	項目1、一番大きい項目、不定長。
財政	項目2、
予算	項目3、
60年度…	項目4、
反響	項目5、
努力は……	見出し、

【表示形式】

通常はA.の見出し1行・書誌事項1行の2行表示、コマンドを打つことでB.の詳細表示モードに切り替わるようにする。

A. 見出し一覧形式（見出しは36字で切り捨て）

1	努力は評価、額には不満 米政府筋示唆__反響
	841230 朝刊 2面 1096頁 5段
2	「予算増額の支援材料に」自民・防衛庁は期待__平和問題研最終報告書提出
	841219 朝刊 2面 704頁 5段

B. 詳細形式（末尾の「見出し」は2行になることあり）

1	努力は評価、額には不満 米政府筋示唆__反響
	841230 朝刊 2面 1096頁 5段
	ID……8412-0687
	項目1…政治
	項目2…財政
	項目3…予算
	項目4…60年度予算政府案
	項目5…反響
	見出し…努力は評価、額には不満 米政府筋示唆__反響
2	「予算増額の支援材料に」自民・防衛庁は期待__平和問題研最終報告書提出
	841219 朝刊 2面 704頁 5段
	ID……8412-0804
	項目1…政治
	項目2…国防
	項目3…<>
	項目4…平和問題研最終報告書提出
	項目5…<>
	見出し…「予算増額の支援材料に」自民・防衛庁は期待__平和問題研最終報告書提出

図2-2-1 データ構造と表示方式

ページまで、1日の夕刊は×ページから×ページまで。縮刷版の各号ごとにこういったテーブルを作成し、和数字の読み変えを行い、見出し文に追加付与した。

こうした手順をへて変換させたデータの一例を図2-2-4に示す。これは後述する簡易システム用のフォーマット構成である。

週間報告

12月28日—1月3日…一七8

1月4日—10日…三六五8

1月11日—17日…六二五8

1月18日—24日…八九五8

座標

生活の豊かさ 都留重人…三二九1

寅さんの家庭論 宮下展夫…六九七1

過熱する生命技術 柴田鉄治…八五九1

国民の意識（本社世論調査）

生活苦高まる声 自衛隊で改憲賛否ほぼ同率…六五7

本社調査に見る国民意識 不満多いが安定を評価…七四1

暮らしと政治・憲法と自衛隊・老後の暮らし・エネルギー源…七四1

仲よくする国…七四9

けなげな女性が理想像 好きな女性は・女性の進出・生まれかわるなら…七五1

どちらの名字・どこまで信用…七五5

日本人らしい平衡感覚 建前と本音をたくみに区別…七五6

調査方法…七五9

{国土・人口}

[人口]

落ち込む出生率 昨年は史上最低 丙午の41年下回る…三1

全国で836万人 年男・年女…三9

人口動態

社会増目立つ中小都市 人口動態 移動率は7%割る…一八七8

世帯構成（55年厚生行政基礎調査）

家庭の老齢化・ミニ化加速 高齢者世帯、242万超す この10年で倍増 …五九一

（表）60歳以上の者のいる世帯の構成割合の推移…五九一2

（表）人員別にみた世帯数と平均世帯人員の動き…五九一4

<開発・保全>

[国土開発]

定住構想

高齢者を念頭に整備「地城づくり」で定住研が報告書…八九二6

[都市計画]

都市計画道路変更案を了承 都計審…一一〇一8

[土地対策]

首都圏の土地利用状況 10メートル平方ごと航空写真に 建設省が新年度着手…六七八1

宅地化へ基礎資料…六七八1

[環境保全]

清流復活北欧へ紹介 環境庁 区民の知恵称賛 江戸川区の親水公園…四六一1

環境保全「ナショナル・トラスト」導入…七七三1

土地買い取りの住民に税金優遇 環境庁方針…七七三1

あくまで住民主体…七七三6

湖沼保全で答申

水辺に憩いの場を 湖沼保全 中公審が答申 市街化に歯止め…九六七1

（解説）汚濁 深刻な状況 省庁の壁超え法整備を…九六九1

湖沼保全の答申内容 要旨…九六九4

<気象>

大寒気団ひかえ81年始動…一三七5

北陸地方にまた寒波…一七〇6

今冬は“準38豪雪” 気象庁が厳戒訴え…二〇九8

なお大雪の恐れ 寒気団がまた接近…二二一9

大雪、平野部は一服 内陸やまず 山形は警報…二五七7

図2-2-2 起こした文書の基本形式

1M	1E	2M	2E	3M	3E	4M	4E	5M	5E	頁データ	01M-05E
1,	0,	0,	0,	65,	0,	89,	0,	113,	137		
6M	6E	7M	7E	8M	8E	9M	9E	10M	10E	頁データ	06M-10E
149,	173,	185,	209,	221,	245,	259,	283,	295,	319		
11M	11E	12M	12E	13M	13E	14M	14E	15M	15E	頁データ	11M-15E
329,	0,	361,	385,	397,	425,	441,	473,	489,	0		
16M	16E	17M	17E	18M	18E	19M	19E	20M	20E	頁データ	16M-20E
513,	537,	553,	577,	589,	0,	621,	645,	657,	681		
21M	21E	22M	22E	23M	23E	24M	24E	25M	25E	頁データ	21M-25E
693,	717,	733,	757,	773,	809,	825,	849,	859,	0		
26M	26E	27M	27E	28M	28E	29M	29E	30M	30E	31M	31E 頁データ 26M-31E
891,	915,	929,	957,	967,	991,	1003,	1027,	1043,	1073,	1085,	1105, 1115

図2-2-3 和数字の発行日への読み替えテーブル

8101-0001 810105 M 5 117 8 <> <> <> 週間報告 <> <> 1 2 月 2 8 日—1 月 3 日__週間報告

8101-0002 810112 M 5 365 8 <> <> <> 週間報告 <> <> 1 月 4 日—1 0 日__週間報告

8101-0003 810119 M 5 625 8 <> <> <> 週間報告 <> <> 1 月 1 1 日—1 7 日__週間報告

8101-0004 810126 M 5 895 8 <> <> <> 週間報告 <> <> 1 月 1 8 日—2 4 日__週間報告

8101-0005 810111 M 1 329 1 <> <> <> 座標 <> <> 生活の豊かさ 都留重人__座標

8101-0006 810121 M 5 697 1 <> <> <> 座標 <> <> 寅さんの家庭論 宮下展夫__座標

8101-0007 810125 M 1 859 1 <> <> <> 座標 <> <> 過熱する生命技術 柴田鉄治__座標

8101-0008 810103 M 1 65 7 <> <> <> 国民の意識(本社世論調査) <> <> 生活苦高まる
声 自衛隊で改憲賛否ほぼ同率__国民の意識(本社世論調査)

8101-0009 810103 M 10 74 1 <> <> <> 国民の意識(本社世論調査) <> <> 本社調査に見
る国民意識 不満多いが安定を評価

8101-0010 810103 M 10 74 1 <> <> <> 国民の意識(本社世論調査) <> <> 暮らしと政治
・憲法と自衛隊・老後の暮らし・エネルギー源

8101-0011 810103 M 10 74 9 <> <> <> 国民の意識(本社世論調査) <> <> 仲よくする国
__国民の意識(本社世論調査)

8101-0012 810103 M 11 75 1 <> <> <> 国民の意識(本社世論調査) <> <> けなげな女性
が理想像 好きな女性は・女性の進出・生まれかわるなら

8101-0013 810103 M 11 75 5 <> <> <> 国民の意識(本社世論調査) <> <> どちらの名字
・どこまで信用__国民の意識(本社世論調査)

8101-0014 810103 M 11 75 6 <> <> <> 国民の意識(本社世論調査) <> <> 日本人らしい
平衡感覚 建前と本音をたくみに区別

8101-0015 810103 M 11 75 9 <> <> <> 国民の意識(本社世論調査) <> <> 調査方法__国
民の意識(本社世論調査)

8101-0016 810101 M 3 3 1 国土・人口 <> 人口 <> <> 落ち込む出生率 昨年は史上最
低 丙午の4 1 年下回る__人口

8101-0017 810101 M 3 3 9 国土・人口 <> 人口 <> <> 全国で8 3 6 万人 年男・年女
__人口

8101-0018 810107 M 3 187 8 国土・人口 <> 人口 人口動態 <> <> 社会増目立つ中小都市
人口動態 移動率は7 %割る__人口動態

8101-0019 810118 M 3 591 1 国土・人口 <> 人口 世帯構成(5 5 年厚生行政基礎調査)
<> <> 家庭の老齢化・ミニ化加速 高齢者世帯、2 4 2 万超す この1 0 年で倍増

8101-0020 810118 M 3 591 2 国土・人口 <> 人口 世帯構成(5 5 年厚生行政基礎調査)
<> <> (表) 6 0 歳以上の者のいる世帯の構成割合の推移

8101-0021 810118 M 3 591 4 国土・人口 <> 人口 世帯構成(5 5 年厚生行政基礎調査)
<> <> (表) 人員別にみた世帯数と平均世帯人員の動き

8101-0022 810126 M 2 892 6 国土・人口 開発・保全 国土開発 定住構想 <> <> 高齢者を
念頭に整備「地域づくり」で定住研が報告書__定住構想

8101-0023 810131 M 17 1101 8 国土・人口 開発・保全 都市計画 <> <> <> 都市計画道路
変更案を了承 都計審__都市計画

8101-0024 810120 M 22 678 1 国土・人口 開発・保全 土地対策 <> <> <> 首都圏の土地
利用状況 1 0 メートル平方ごと航空写真に 建設省が新年度着手

8101-0025 810114 M 21 461 1 国土・人口 開発・保全 環境保全 <> <> <> 清流復活北欧
へ紹介 環境庁 区民の知恵称賛 江戸川区の親水公園

8101-0026 810123 M 1 773 1 国土・人口 開発・保全 環境保全 <> <> <> 環境保全「ナシ
ョナル・トラスト」導入__環境保全

8101-0027 810123 M 1 773 1 国土・人口 開発・保全 環境保全 <> <> <> 土地買い取りの
住民に税金優遇 環境庁方針__環境保全

8101-0028 810123 M 1 773 6 国土・人口 開発・保全 環境保全 <> <> <> あくまで住民主
体__環境保全

8101-0029 810128 M 1 967 1 国土・人口 開発・保全 環境保全 湖沼保全で答申 <> <> 水
辺に憩いの場を 湖沼保全 中公審が答申 市街化に歯止め

8101-0030 810128 M 3 969 1 国土・人口 開発・保全 環境保全 湖沼保全で答申 <> <> (
解説) 汚濁 深刻な状況 省庁の壁超え法整備を__湖沼保全で答申

8101-0031 810128 M 3 969 4 国土・人口 開発・保全 環境保全 湖沼保全で答申 <> <> 湖
沼保全の答申内容 要旨__湖沼保全で答申

図2-2-4 変換後、日付、面名などがはいったデータ

3. データづくり

3.1 データ起こし

電子データがない過去の文献データをどうやって起こすかということは、これから新しくデータベースをつくらうというところにとって、最大の関門である。

とくに縮刷版データベースは戦後編だけでも対象期間が50年近くにもなろうというものであり、その入力費用は膨大なものになることが予測された。

一番簡単な方法はワープロ外注である。しかし、単純試算してみると、1カ月分のデータは約16万文字あった。1字1円としても1カ月で16万円、年間にすれば200万円近く、10年分になればこれだけで2,000万円近いものになってしまう。

3.2 光学読み取り機の導入

そこで、実験的に導入したのが光学的読み取り機（OCR）である。

新聞社では、原稿の入力の省力化が長年の課題となってきた。ワープロが普及する以前は専門の入力要員を多数本社や支局に配置し、原稿の送稿、入力の迅速化を図ってきた。

ワープロが普及した現在、記者が自分で原稿をワープロでたたき、電話線を通じて送稿するというのは一般化してきたが、重大裁判の判決のようなときは、人海戦術で入力することにはかわりはない。

そういった時にOCRが戦力になるのではないか、新しいデータベースを起こすとき利用がきかないか、などの発想でOCRをテスト的に導入していたのをこちらにふりむけることにした。

当初は富士電機製のXP-50を使っていたが、途中から東芝製のExpressReaderJ70という機種に変更した。

XP-50はホストマシンにNECのパソコンPC9800型が使えるというのが魅力であった。また導入当時（90年8月）では、製品化していた数社のものと比較検討してもっとも優れた性能をもっていた。カタログ値では文字認識速度30文字／秒、文字認識率99%ということであった。

ExpressReader（92年2月導入）は東芝のJ3100ZDがホストマシン。新しい機種だけに文字認識速度70-100字／秒、認識率99.5%以上（いずれもカタログ値）と読み取り速度、認識精度が大幅に向上していた。

OCRとは、入力したい文献をスキャナーにかけ、その画像データを分析し、文字を推定し、ASCIIコードとして出力するというものである。

スキャナーにかけた文献はまず走査密度に応じた細かい点（ドット）に分解される。走査密度が細かくすれば、分解性能が上がり解析能力も上がりそうなものだが、解析処理速度とのかねあいもあるらしく、400DPI（一インチあたり400本の走査線）のものが主流となっている。

文字の認識方法は各社各様であるが、一般的にはまず、読み込んだ白黒の点のかたまり（画像）から、写真やイラストをより分けて文字部分だけを取り出す（これは手動で指定することも多い）。次にこの文字部分の画像の中から行間、字間の区切り（空白）をもとに文字画像のかたまりを1字ずつ切り出す。

それから、OCR内蔵の解析機があらかじめもっている文字の母型パターンと照合して、ある程度の共通点がみられたら文字コードを決定する、あるいは文字画像の縦線、横線、ハネ具合などの方向から文字を推定しコードを発生させるなどの方法がある。

しかし、OCRは読み取り精度が100%というわけにはいかず、読み取りミスをはたにチェックするかが大きな課題であった。

XP-50は「21」「35」といった縦書きの連数字がまったく読めず、読み取り不能のゲタ字（＝）として出力された。

また「輪」と「輪」、「土」と「土」など似たような形の書体は読み取りミスを起こしやすく、平均的読み取り精度は92-93%くらいであった。

特定の文字をいつもきちんと読み違えるのならば、それなりの対処のしようもある。ところが、OCRでは読み取りスキャナーに原稿をかける微妙な角度の違いで、読み取り結果に微妙な差異が出る。同一ページ内にある同じ文字でも、正しく読めているものとそうでないものがあるといった具合である。

強制修正をかけた方がよいか、それとも修正しない方が他への波及被害が少ないかなどを体験的に積み重ね、最終的には図3-2-1のような300字くらいの読み変えテーブルを作成した。

OCRで読ませたデータはその場で修正ができる。文字修正モードにし、怪しい箇所あるいは明らかに読み誤っている箇所にカーソルを置き、マウスをクリックすれば、他の候補文字がいくつか表示される。該当する文字があれば、その文字をクリックすれば正しい文字に置き換えられるという仕組みである。

—Yn—
 =Yn
 —Yn
 =Yn
 Yn—
 =Yn
 —□
 =言
 —フ
 =ラ
 —フ
 =ラ
 エハ
 =六
 エ合
 =六
 エハ
 =六
 兀
 =元
 —向
 =高
 —兀
 =元
 旧本
 =日本
 貝
 =貝
 峙
 =時
 Yn肉
 =Yn [
 、つ
 =づ
 、ハ
 =パ
 束
 =束
 約束
 =約束
 拘束
 =拘束
 束縛
 =束縛
 結束
 =結束
 里—
 =黒
 竺二
 =三
 エご
 =六一
 柑
 =相

図3-2-1 読み間違えた文字の変換テーブル

XP-50は、読み取り不能の文字は「=」あるいは特殊な記号文字で表示された。

ExpressReaderでは、処理が進歩しており、マシンが読み取りに自信がないものについては、通常白地に黒文字表示のところは赤字で表示される。さらにその横に読み取った原稿のスキャン画像が表示され、いちいち原稿を照合することなしに、校閲作業ができることになっていた。(XP-50の改良型、XP-70では原稿の画像表示ができるように改良された)。

ところがExpressReaderで問題なのは、赤字表示部分だけが読み損ね、あるいは読み損ね候補であるかといえはそうではなく、マシンが問題なしと判断した箇所でも結構読み損ねがあるということであった。メートル、キロなどといった連字にこの傾向が強かった。

当初はOCR入力したその場で完全な校正までできるのではないかという期待を抱いていたのであったが、こういった事情では、やはり全文校閲をする必要が生じる。

いろいろ作業性などを考えてテストしてみたが、最終的にはOCR入力担当者は赤字表示部分だけを直す、その上で初校ずみデータをMS-DOSファイルに落とし、これを原本とともに再校閲に出すという二重チェックの方法をとった。

オペレーターの養成は容易である。XP-50は構造が単純なだけ、操作も簡単で、パソコンにさわるのが初めてと言う学生アルバイトでも数時間の研修で一人立ちさせることができた。

1日6時間の作業時間で平均14-16万字、縮刷版にして約1カ月分の入力できた。熟練したワープロ入力者が1日に打てる文字量は1万数千字というから、不完全とはいえ、素人が数人分の仕事をこなすという計算になる。

問題は直しであるが、新聞社では書いた原稿には誤字はつきもの、それはまた専門の校閲スタッフがチェックするという分業体制を取っている。その考え方でいけば、少々の間違ひはつきもの、どっちみち校閲を通すのであるから、それが7-8%であろうと、それ以下であろうとあまり差はないように思われた。

正読率93%ということは、100字読ませたうち93字は正しく読んだということの意味する。素人がワープロを打つより精度がよいかもしれない。しかし、実際の作業としてはなかなか負担が大きかった。

直しを入れた結果の一例を図3-2-2に示す。

ワープロ入力業者の人とOCRの利用について話し合ったことがあるが「正読

← 臨床試験 厚生省研究班七五五

(解説) 不純物排除がカ

← ギ……七五五 インターフェロン、米社から供給契約 ミドリ十字……六八八八

生薬ブームご用心 厚生省モニター報告 漢方薬にも副作用……六八三五

「ワクチン」郵送開始……八〇六二

[衛]生

無人、六店に五店 コイン

← ランドリー 都衛生局の実態調査……三六九 来年早々にも細菌検査……三六九一

炭酸飲料に殺菌義務づけ 食用缶にも初の品質規格 食品衛生調査会答申……六四七1

ケーキの二割、衛生上「不良」食品一斉取り締ま

← り……九一〇七

← つめ

← 福祉・厚生

← つめ

父子家庭にやっとう 援軍ヘルパーや介護人 都審議会提言受け準備……一〇一六

← 二輪のサンタ 千台が集合 恵まれぬ子にオモチャ 江の島……五二五

朝日福祉設備助成金 都内は四団体……六〇七1 タイプにかける 墨田区肢体不自由児父母の会半八三

← 七1とけ合う心実現 重度肢体障害者共同生活寮……六二七1

← 四団体に朝日福祉設備助成金を贈呈(都関係)……八三八八

保育科「勝手な値上げ違法」 渋谷区長らを訴え

る……九〇四——

ハンセン病患者の悲願……年

← 「人間回復の橋」に予算 瀬戸内海の長島……九四六1

福祉施設建設 賛否の住民投票反対 埼玉・鳩山の自治会内 排斥の動き……九七八5

「けやきの郷」来月日に投票実施 自治会がビ

← ラ配る……一〇二八9

社会保険

社会保険

老人保健法案 年7700円も負担増 健保連試算拠出金は憶外大幅……五四一3

[老人問題]

老人福祉「買う」時代 一般世帯へもヘルパー 所得に応じ負担……三九一1 社福審提言……三九一1

(解説) 老人大国へ福祉も備え 中流層に心くばり 社福審の提言……三九二1

老いまた楽し 我らがママよ、ありがと 昔の常連喜寿祝い……四五三1

この働きバチ列島「いつまでも」 歳定年、76%が賛成……七四九1

図3-2-2 ハードコピーに朱を入れた例

率97%以上なら利用価値がある。それ以下だったら熟練したパンチャーの打つ方が効率的」という答であった。ほぼ同感である。

XP-50は連数字が読めないことがネックとなり、実績として縮刷版では93%までしか正読率が上がらなかった。横書き文書では半角数字でも読め、正読率は97%くらいまであげることができた。

一方、ExpressReaderの方は、修正作業の環境に進歩が見られたほか、ホストマシンが強力になっている分だけ、処理速度の向上がみられた。読み取り精度も大幅に向上しており、実績平均99%以上の正読率であった。

ところが、こちらにもそれなりの弱点があった。

OCRでは一般的にまず読み取らせたい文献をスキャナーにかけ、表示される画像のうち文字化したい領域をマウスで指定し、その部分を解析にかけ文字に変換させる。

XP-50ではこの作業が完全手動であったが、ExpressReaderではレイアウト自動解析といってスキャンした画像を自動解析し、写真やグラフ部分を除去し、文字部分を自動的に領域指定するという機能がついていた。

だが、縮刷版索引は、段ごとに仕切罫が入っている比較的単純なレイアウトであるにもかかわらず、これが結構難渋した。2段を1段にまとめてしまったり、逆に1段の中をさらに細かく分けてしまうなどの事態が発生した(図3-2-3)。

領域指定に失敗したので、再解析したらまた同じ間違いを繰り返す、結局始めから手動でやり直すという作業を強いられた。ExpressReaderでは手動による部分的手直しがきかず、自動モードにすると完全に機械任せ、手動モードにするとすべてが手動という極端な操作法が効率を著しく低下させた。

もう一つの問題は文字認識の精度は高いが、行頭1字下げ、2字下げという空白認識が甘いということであった。縮刷版索引では、見出し本文は1行でおさまることは少なく、2行、3行にわたっている場合が多い。通常1行目は行頭1字下げ、2行目以降は2字下げという体裁になっている。

これを1本のレコードにするためには、行頭2字下げの部分はスペースと前の行末の改行記号を削除し、前の行につなぐという作業が必要になる。この行頭空白の認識がでたらめだと、自動的処理が困難になる。

この点、XP-50は空白スペースを正しく識別したが、ExpressReaderの方は画像自動解析を取り入れているせいか、この点がかなり弱かった。極端な場合は、本

図3-2-3 レイアウト解析に失敗した例

来頭下げになる部分が突き上げになり、逆に突き上げになるべき部分が数文字分頭下げになるなど、常識では首をひねりたくなるケースも散見された。手動で体裁直しを行うため、かなりのブレーキになった(図3-4)。

結局、OCR入力というのは、スキニング、文字領域の確定、解析、さらにその場での校閲をつけ加えるなら、これらを総合したトータルなものである。

解析速度だけが早くても、スキニングや文字領域の確定のところで手間取ったのでは効果は半減である。校閲をぬきにした場合、スキニングから領域確定までの時間と解析にかかる時間との割合はXP-50が3:7、うまくいった場合のExpressReaderが5:5といったところであろう。その場での校閲作業を加えると、さらに解析時間の占める比率は小さくなる。

縮刷版見出しデータベースで、入力と校閲を分離したのは、一つはOCR本来の目的である文字入力にアイドルタイムを生じさせたくなかったからであった。OCRがパソコン並みの価格になれば、それなりの用途があろうが、1台数百万円というような現状では、入力に専念させる方が有効利用に結びつく。

実績として、OCRによる入力は、機器本体の購入費用を無視すれば、ワープロ外注に比べ30%から40%のコストでできる感じであった。

最近、多くのメーカーがOCRに参入し、製品も発表されている。導入にあたっては、目的(どこまでの精度の文献をめざすかを含めて)、文献の形態(単純なレイアウトか、新聞切り抜きのような複雑なものか等)、最終形態までの作業手順をどうするかなどを考慮した上で、その要求に見合った機種を決定すべきであろう。

への提言」 入選者決ま る……………三五—4	への提言」 入選者決ま る……………三五—4
〔解説〕 燃えぬ日本の反核 デモ 危機感の薄さ反 映……………四三六—1	〔解説〕 燃えぬ日本の反核 デモ 危機感の薄さ反 映……………四三六—1
日本含む非核地帯を 学者 ら三千六百人声明……………五九四—1	日本含む非核地帯を 学者 ら三千六百人声明……………五九四—1
バルメ委東京会議 「米ソの軍縮交渉注目」 バ ルメ委東京会議 委員長 来日し語る……………九九七	バルメ委東京会議 「米ソの軍縮交渉注目」 バ ルメ委東京会議 委員長 来日し語る……………九九七
軍縮運動に被爆者の声 バ ルメ委きょう開幕……………一二—1	軍縮運動に被爆者の声 バ ルメ委きょう開幕……………一二—1
「非核」推進で一致 飛鳥田 ・バルメ会談……………一二—8	「非核」推進で一致 飛鳥田 ・バルメ会談……………一二—8
平和と軍縮めざして 「パ ルメ委員会」が開会……………一四三—6	平和と軍縮めざして 「パ ルメ委員会」が開会……………一四三—6
「日本も軍縮に努力」 首相、 バルメ氏と会談……………一六五—8	「日本も軍縮に努力」 首相、 バルメ氏と会談……………一六五—8
軍拡競争やめなさい パル メ氏 歓迎宴で説く……………二〇四—1	軍拡競争やめなさい パル メ氏 歓迎宴で説く……………二〇四—1
戦略核も削減を 米ソに交 渉呼びかけ バルメ委二三五—5	戦略核も削減を 米ソに交 渉呼びかけ バルメ委二三五—5
いま国際的安保を 研究集 会でバルメ提言……………二五九—1	いま国際的安保を 研究集 会でバルメ提言……………二五九—1
経済力で平和に貢献 バル メ委が日本に要望……………二七五—1	経済力で平和に貢献 バル メ委が日本に要望……………二七五—1
軍縮願ひ活発な討議 バル メ委関連の研究集会か ら……………二七六—1	軍縮願ひ活発な討議 バル メ委関連の研究集会か ら……………二七六—1
各地域に非核地帯を 育 てよう草の根運動……………二七六—1	各地域に非核地帯を 育 てよう草の根運動……………二七六—1
世界に届け ヒロシマの 声 軍縮研究集会を間い 一、今堀誠ニ……………二七六—6	世界に届け ヒロシマの 声 軍縮研究集会を間い 一、今堀誠ニ……………二七六—6
世界平和へ大きな役割 バルメ氏、日本に期待三五—9	世界平和へ大きな役割 バルメ氏、日本に期待三五—9
▽広島集会 広島でも研究集会……………一二—3	▽広島集会 広島でも研究集会……………一二—3
「反核、被爆者に学んで」 バルメ委あす広島対話集 会……………二五九—1	「反核、被爆者に学んで」 バルメ委あす広島対話集 会……………二五九—1
広島に着く バルメ委メン バー……………二七五—7	広島に着く バルメ委メン バー……………二七五—7
バルメ委、全日程終える 「ヒロシマ」に衝撃 市民	バルメ委、全日程終える 「ヒロシマ」に衝撃 市民

(OCRの読んだ例)

への提言」 入選者決ま る……………三五—4	への提言」 入選者決ま る……………三五—4
〔解説〕 燃えぬ日本の反核 デモ 危機感の薄さ反 映……………四三六—1	〔解説〕 燃えぬ日本の反核 デモ 危機感の薄さ反 映……………四三六—1
日本含む非核地帯を 学者 ら三千六百人声明……………五九四—1	日本含む非核地帯を 学者 ら三千六百人声明……………五九四—1
バルメ委東京会議 「米ソの軍縮交渉注目」 バ ルメ委東京会議 委員長 来日し語る……………九九七	バルメ委東京会議 「米ソの軍縮交渉注目」 バ ルメ委東京会議 委員長 来日し語る……………九九七
軍縮運動に被爆者の声 バ ルメ委きょう開幕……………一二—1	軍縮運動に被爆者の声 バ ルメ委きょう開幕……………一二—1
「非核」推進で一致 飛鳥田 ・バルメ会談……………一二—8	「非核」推進で一致 飛鳥田 ・バルメ会談……………一二—8
平和と軍縮めざして 「パ ルメ委員会」が開会……………一四三—6	平和と軍縮めざして 「パ ルメ委員会」が開会……………一四三—6
「日本も軍縮に努力」 首相、 バルメ氏と会談……………一六五—8	「日本も軍縮に努力」 首相、 バルメ氏と会談……………一六五—8
軍拡競争やめなさい パル メ氏 歓迎宴で説く……………二〇四—1	軍拡競争やめなさい パル メ氏 歓迎宴で説く……………二〇四—1
戦略核も削減を 米ソに交 渉呼びかけ バルメ委二三五—5	戦略核も削減を 米ソに交 渉呼びかけ バルメ委二三五—5
いま国際的安保を 研究集 会でバルメ提言……………二五九—1	いま国際的安保を 研究集 会でバルメ提言……………二五九—1
経済力で平和に貢献 バル メ委が日本に要望……………二七五—1	経済力で平和に貢献 バル メ委が日本に要望……………二七五—1
軍縮願ひ活発な討議 バル メ委関連の研究集会か ら……………二七六—1	軍縮願ひ活発な討議 バル メ委関連の研究集会か ら……………二七六—1
各地域に非核地帯を 育 てよう草の根運動……………二七六—1	各地域に非核地帯を 育 てよう草の根運動……………二七六—1
世界に届け ヒロシマの 声 軍縮研究集会を間い 一、今堀誠ニ……………二七六—6	世界に届け ヒロシマの 声 軍縮研究集会を間い 一、今堀誠ニ……………二七六—6
世界平和へ大きな役割 バルメ氏、日本に期待三五—9	世界平和へ大きな役割 バルメ氏、日本に期待三五—9
▽広島集会 広島でも研究集会……………一二—3	▽広島集会 広島でも研究集会……………一二—3
「反核、被爆者に学んで」 バルメ委あす広島対話集 会……………二五九—1	「反核、被爆者に学んで」 バルメ委あす広島対話集 会……………二五九—1
広島に着く バルメ委メン バー……………二七五—7	広島に着く バルメ委メン バー……………二七五—7
バルメ委、全日程終える 「ヒロシマ」に衝撃 市民	バルメ委、全日程終える 「ヒロシマ」に衝撃 市民

(空白を手動で正しく修正後)

への提言」 入選者決ま る……………三五—4	への提言」 入選者決ま る……………三五—4
〔解説〕 燃えぬ日本の反核 デモ 危機感の薄さ反 映……………四三六—1	〔解説〕 燃えぬ日本の反核 デモ 危機感の薄さ反 映……………四三六—1
日本含む非核地帯を 学者 ら三千六百人声明……………五九四—1	日本含む非核地帯を 学者 ら三千六百人声明……………五九四—1
バルメ委東京会議 「米ソの軍縮交渉注目」 バ ルメ委東京会議 委員長 来日し語る……………九九七	バルメ委東京会議 「米ソの軍縮交渉注目」 バ ルメ委東京会議 委員長 来日し語る……………九九七
軍縮運動に被爆者の声 バ ルメ委きょう開幕……………一二—1	軍縮運動に被爆者の声 バ ルメ委きょう開幕……………一二—1
「非核」推進で一致 飛鳥田 ・バルメ会談……………一二—8	「非核」推進で一致 飛鳥田 ・バルメ会談……………一二—8
平和と軍縮めざして 「パ ルメ委員会」が開会……………一四三—6	平和と軍縮めざして 「パ ルメ委員会」が開会……………一四三—6
「日本も軍縮に努力」 首相、 バルメ氏と会談……………一六五—8	「日本も軍縮に努力」 首相、 バルメ氏と会談……………一六五—8
軍拡競争やめなさい パル メ氏 歓迎宴で説く……………二〇四—1	軍拡競争やめなさい パル メ氏 歓迎宴で説く……………二〇四—1
戦略核も削減を 米ソに交 渉呼びかけ バルメ委二三五—5	戦略核も削減を 米ソに交 渉呼びかけ バルメ委二三五—5
いま国際的安保を 研究集 会でバルメ提言……………二五九—1	いま国際的安保を 研究集 会でバルメ提言……………二五九—1
経済力で平和に貢献 バル メ委が日本に要望……………二七五—1	経済力で平和に貢献 バル メ委が日本に要望……………二七五—1
軍縮願ひ活発な討議 バル メ委関連の研究集会か ら……………二七六—1	軍縮願ひ活発な討議 バル メ委関連の研究集会か ら……………二七六—1
各地域に非核地帯を 育 てよう草の根運動……………二七六—1	各地域に非核地帯を 育 てよう草の根運動……………二七六—1
世界に届け ヒロシマの 声 軍縮研究集会を間い 一、今堀誠ニ……………二七六—6	世界に届け ヒロシマの 声 軍縮研究集会を間い 一、今堀誠ニ……………二七六—6
世界平和へ大きな役割 バルメ氏、日本に期待三五—9	世界平和へ大きな役割 バルメ氏、日本に期待三五—9
▽広島集会 広島でも研究集会……………一二—3	▽広島集会 広島でも研究集会……………一二—3
「反核、被爆者に学んで」 バルメ委あす広島対話集 会……………二五九—1	「反核、被爆者に学んで」 バルメ委あす広島対話集 会……………二五九—1
広島に着く バルメ委メン バー……………二七五—7	広島に着く バルメ委メン バー……………二七五—7
バルメ委、全日程終える 「ヒロシマ」に衝撃 市民	バルメ委、全日程終える 「ヒロシマ」に衝撃 市民

図3-2-4 行頭の空白認識に失敗した例

3.3 校正作業

次に校正とデータ修正の問題である。

OCRの利用は前提のようなものであったが、校正作業にもいくつかの選択肢があった。プロの校閲に任せることができるならこれにこしたことはない。しかし、これまた1文字1円とかいう相場であり、それではなんのためのOCR入力かということになってしまう。

そこでコスト切り詰めを図るため、家庭在宅婦人労働力の活用を考えた。子育てなどで家庭からは離れられないものの、在宅でできる仕事ならやりたいという高学歴の女性が増えているのではないかと考えたわけである。納期の期限を切った仕事ならそれに見合うものを払わねばならないが、期限を緩くすれば、それなりの単価で仕事を引き受けてもらえるのではないかという発想であった。ワープロやパソコンがかなり家庭に普及しているのではないかという期待もあった。

ニューメディア本部に勤務していた退職者、あるいは知人のつてなどを頼りに10人くらいを募集、熟練者なら時給1500円くらいになるように作業量を設定して仕事を委託してみた。

初心者がほとんどだったので、校正の記号、約束ごとからマニュアルが必要であった。しかし、何度かやりとりするうちにマニュアルもおのずと確立し、一応それなりにはやれるという結果は得た。

しかし、こういった直取引の作業委託はその管理が大変であった。パソコン、プリンターを所持している家庭は予期したより少なかった。そこで、校正とフロッピーの直し作業を分離した。

まずこちらでプリントアウトを行い、それを原本とともに荷造りして宅急便で校正者に発送する。校正が終わり、返送されてきた朱入りプリントアウトを今度はフロッピーとともにパソコンをもっている家庭に発送する。月末には作業量を集計して精算する。これらの作業だけで一人の専従要員を必要としかねぬ形勢になってきた。

次に考えたのが在宅婦人を組織しているパソコンネットとの提携であった。

私鉄系の会社の組織している半商用ネットでこういった在宅の仕事を請け負うところがあり、試験的に作業を委託してみた。

しかし、一括委託にもそれなりの問題があった。

会員のレベルにかなりのばらつきがあり、出来、不出来にかなりの差があった。

直接のやりとりでないため、書式などこちらの要望がうまく伝わらない、校正ミスが多発した場合の責任の所在が明らかでない、などである。請け負い元を通すことになるので、報酬に対する不満が末端の作業者にはあるようにも思われた。

ここでも当初、校正とフロッピー直しを分離したが、校正のミスをそのまま見過ごすケースが続出したため、対策を迫られた。

何度かネットの窓口と話し合いを持ち、作業担当者を固定、担当者名も明らかにし、責任の所在を明確にすることにした。半日間の出張研修を行い、発注側の意図、作業手順を説明した。さらにネット内に作業連絡ボード（電子掲示板）を設け、疑問点や要望などをパソコン通信で担当者同士がやりとりするという仕組みをつくり、ようやくほぼ満足できるレベルに達することができた。

校正がかえってきたら最終チェックである。上位項目の見出しを自動的に付加するためには、書式の約束ごとがきちんと守られていなければならない。ところが長年のうちに縮刷版本体の表記が手直しされていることがあった。こんな時はパソコン通信のボードが役立ったが、細かいところでは手の回らないところも出てくる。

作業委託にあたっては「原稿通り」というのを原則にした。10ページくらいのマニュアルを用意したが、あまり約束ごとが多いと作業効率が落ちることを懸念したためだ。

また、OCRの癖として和数字の「一」「二」「三」の識別があいまいであった。「一二三」「二二二」「三二一」は縦書きの場合、区別がつきにくいと見えてしばしば読み損ねがあった。初期の段階では校正者も不慣れだったため、ここの見落としが頻発した。この和数字部分の見落としは日付、掲載面の手がかりになるだけに影響が大きいので、気を遣った。

4. システム構築

4.1 簡易システム構築

縮刷版データベースは商用オンラインに載せないと利用者が限定されるので、最終的には朝日新聞記事データベースと同じ方式でデータベースを構築、一般提供をめざして作業が進行中である。

しかし、この方式は時間がかかる。手軽に全体像がつかめるシステムがほしい。

そこで、本格的編集を進める一方、当社が松下電器産業株式会社東京情報システム研究所と共同研究で開発を進めている新しい方式の情報検索システム「検蔵君」で、この縮刷版データベースの簡易版を構築した。

一般に日本語文献のデータベースは、日本で最大規模の日経テレコンを含め、文献から適切な索引語（キーワード）を切り出し、あるいは付与して、その索引語ファイルに検索をかけ、そこから目的の文献をひきだしてくるという方式が主流である。

全文書を頭から終わりまでサーチするより、索引語ファイルを探した方が、余分な情報も少なく、また検索しやすい構造になっているので大量、高速の検索が可能になるためである。

こういった検索システムを築くためには、KW切り出しや検索のために大型汎用コンピューターを使った大規模なシステムが必要であった。

しかし、最近の電子技術の進展、とくにワークステーションの高性能化にともなって、こういった索引語ファイルを別に作成するのではなく、テキスト全体をそのまま始めから終わりまでサーチし、該当する文字列を含む文書を拾いだし、データベースとして機能させるというフルテキストサーチ型データベースが現れてきた。

この方式は個々のキーワードの切り出しが不要のため、省人化が可能で、手軽に容易にデータベースが構築できる、環境次第でかなりのところまで自動化ができるという特徴がある。

もっとも収録文書100万件という大規模データベースにとってかわるにはパワー不足が否めないが、今後の技術動向次第では、十分処理能力をもつものに育つ可能性を秘めている。

当社が新聞記事データベース構築に着手した1984年ごろは、こういった低価

格・高性能のマシンは存在しておらず、また文字列のフルサーチなどは効率の面からとても実用に耐えるものとは思われていなかった。

その後、ソフト・ハード技術の進歩で、こういった方式のデータベースが実用化の段階に進んできたわけである。

4.2 全文フルサーチ型検索システム

「検蔵君」はUNIXマシンを検索サーバーとし、VMEバス仕様の高速検索エンジンを搭載したシステムである。検索エンジンは文字列検索専用LSI（ストリームデータプロセッサSDP）をVMEバス3枚の基盤組み込んだもので、ここにメモリーに読み込んだデータを一気に流し、解析を行う。

検索サーバー機としては、Solbourne（ソルボーン）、PanaStation、SunのSPARCステーションの一部などが可能だが、当社ではSolbourne5/600型（メモリー32MB、ハードディスク660MB）を使用している。

この検索システムの特徴としては

イ）データベースの構築・保守が簡単

ワープロ、OCRなどで作成したテキストデータをそのまま使える
キーワードづけ、索引ファイルの作成などの前処理や定期保守が不要

ロ）使いやすく正確な検索

自由な文や単語で検索できる
同義語辞書を使って検索漏れを減らす
単語認識で検索ノイズを減らす

ハ）検索専用ハードウェアを使った高速検索

同義語を使って複雑になった条件でも検索速度が低下しない
検索速度：約10MB（500万文字）／秒

ニ）ネットワーク対応

1台のサーバーを複数のクライアントが共同利用。
分散データベースと多彩なマンーマシン－インターフェイスを実現
できる

などがあげられる。

ワープロやパソコンの急激な普及で家庭や事業所に大量の電子情報が蓄積され

ている。しかし、その情報も一度報告書などに使われてしまうと、大半はフロッピーのなかで半ば「死蔵」されている状態である。

これらに文書を一定の書式につないでいけば、そのままデータベースにできるというところに「検蔵君」の一つの特色がある。

パソコンソフトの中にも、こういったフルサーチ方式で文献検索を行うデータベースソフトも現れてきているが、処理量、速度の問題でまだ実用にはほど遠いレベルである。

また、キーワード切り出し不要ということは、そのための要員が大幅に省略できるという利点をもつ。

一方、キーワードなし検索という検索はこれまで出来なかった新しい用途も生み出した。索引ファイルによる検索は当然、ある意味をもったキーワードしか検索対象としないが、文字列一致だと、動詞や形容詞、さらには「を、」といった従来ではまったく考慮の対象外であった文字列まで検索することができる。これは言語の用法研究などには大きな助けとなることであろう。

この「検蔵君」で81年から84年分までの見出しデータベースをとりあえず構築した。10年分としなかったのは、一つはハードディスクの容量の関係、もう一つは朝日新聞記事データベースが85年から本格編集体制に入っており、緊急性が薄かったためである。

縮刷版見出しデータベースの元データ作成には一貫してパソコン（NECのPC9800）を使用した。パソコンやワープロの文書から簡単にデータベースを構築する道を探るというのがひとつの目的だったからである。

校正のすんだ文書をパソコンのエディターで1カ月単位にまとめ、変換ソフトで大・中・小項目の見出しを各見出し文の頭に付加、さらに日付、面名などをつけ加えた。

このデータをイーサネット経由でワークステーションに転送し、シフトJISコードからEUCに変換、「検蔵君」データファイルに変換した。

「検蔵君」ソフトはユーザーが変更可能な項目テーブルをもっており、入力するデータの書式に応じてテーブルの書換を行えば、読み込んだデータがそのままデータベース化できる。

本業の合間を縫っての作業であったため、1カ月近くを要したが、データが完成してから実際にデータベース構築に要した時間は実質的には数時間であった。

<縮刷版見出しデータベース 簡易システムの流れ>

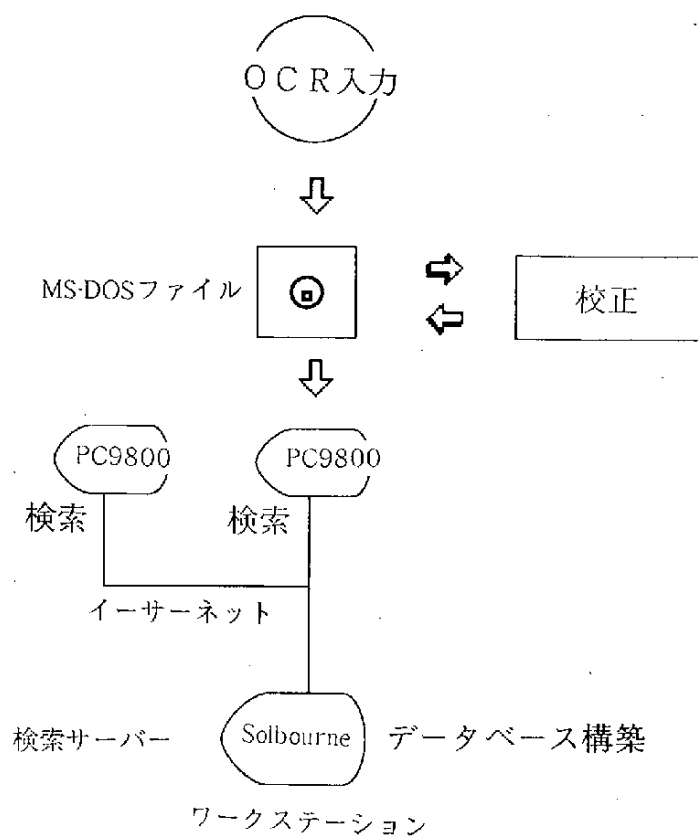


図4-2 縮刷版見出しデータベース簡易システムの流れ

4.3 H I A S K 編集システムの改造

H I A S K 編集システムは当初、新聞記事をデータベース化することを目的に設計されており、ネルソン（朝日新聞の新聞組版システム）以外のシステムで制作されたテキストを取り入れる機能がなかった。

編集作業の端末システムも、全端末が直接ホストと直結し、おのおのがホストの中に作業領域を占有する T S O（Time Sharing System）方式である。

端末の機能が限定されていた10年前はともかく、パソコンの機能向上に伴い、端末でできる仕事は端末で、ホストでしか出来ないことはホストでという、いわゆるダウンサイジングの必要性が痛感されるようになってきた。

これらの事情から1992年3月から12月にかけて、編集システムの一部改造を行った。

システム改造の主な目的は

- (1) 現行システムを生かしたままの編集の合理化
- (2) 多様なテキストからデータベース構築の道を開く
- (3) 分散処理指向における試験的なシステム化

である。

一挙に新システムを目指すことは予算的にも、システムの的にも無理なので、現行の編集方式と並存できる形を目指した。

概略は図4-3の通りだが、ホストサーバーはC I C S トランザクションとし、現行のデータベース編集環境との接続インターフェイスをもたせた。

ホストサーバーとL A Nサーバーの接続はS N Aセッションタイプ6.2を用いた。

L A Nサーバー側での送信処理に関してはA P P Cを使い、3270ファイル転送ではできない、機能上の弾力性をもたせた構成や、対話環境の構築を目指した。

L A Nサーバーとクライアント端末間の接続はトークンリングネットワークによる接続とし、作業において全クライアント（8台）が直接L A Nサーバーのファイルにアクセスできるようにした。

L A NサーバーのO SはO S / 2、端末はD O S / VをO Sとし、端末の編集には市販のV Zエディターを採用した。新たにエディターを開発するより、市販のものにマクロ処理を加えた方が工程的にも安上がりだし、バージョンアップにも対応しやすいと考えたからである。

このシステムを平たくいえば、まず夜のうちに記事ばらしされたテキストはホ

ストのファイルに蓄積される。これは従来通りだが、漢字カナ変換を通さない、文法情報をもたないただのテキストであるところが違う。

このテキストを各端末がダウンロードし、書誌的事項の手直し、見出し直し、分類語付加などを行う。これは普通のテキスト文書だから、編集に使うツールはワープロでもエディターでも何でもよい。

KW切り出しは端末ではできないから、一度ここで直したテキストをホストにアップロードし、KW切り出しをおこなう。これはアップロードと同時に自動的に行われる。

KW切り出し結果を再びダウンロードし、プリントアウトしデスクチェック、読み損ね、編集ミスなどは直接KW文書を直す。直しが終わると再び、アップロード、ホスト側で従来のKW文書のフォーマットに変換され、外部送信用ファイルと蓄積用ファイルに送られる。

エディターによる端末編集の利点は、端末の機能を駆使した機動的運用が行えるということである。

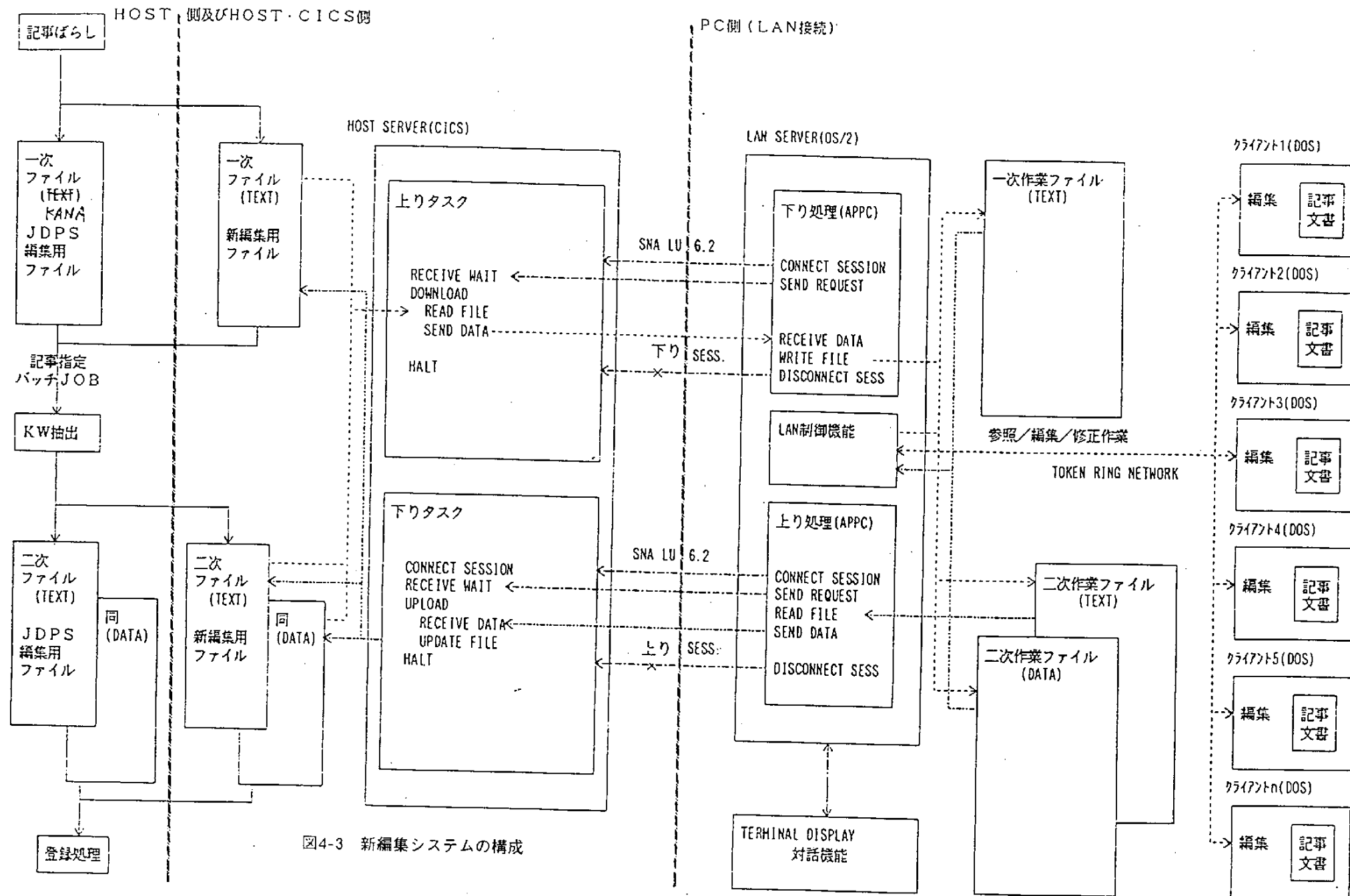
T S O編集では、端末側でコマンドを一つ打つたびに、それがホスト側に伝えられ、ホストがそれを処理し、結果が返ってくる。

つまり改行ひとつ打つにもコマンドがホストに伝えられ、ホストを動かすわけである。この結果、改行ひとつに数秒間を要するという状況になる。

また大型ホストでは、パソコンやワープロではほぼ常識の機能である、削除した部分の復活、特定の部分のブロック移動などが不可能とされていた。

パソコン機能の導入でこういった編集上の制約から解放されたほか、ホスト内に受け入れ環境さえ整えば、フロッピーによるテキスト入力も可能という道が開けた。

スタンドアローン型編集を採用するにあたっては、分類語の自動変換やKW文書を直接メンテすることによるフォーマット崩れの防止などいくつかの問題点があったが、V Zのマクロ機能を駆使することにより、現行の機能はほとんどカバーしたシステムとすることができた。



5. 検証

5.1 プロトタイプでの検索例

「検蔵君」で構築した縮刷版見出しデータベースを使った検索例を次に紹介する。

検索画面の構成は次のようになっている。

まず文書選択画面になる。

＜文 書 選 択＞

朝日新聞見出し1981年
朝日新聞見出し1982年
朝日新聞見出し1983年
朝日新聞見出し1984年
朝日新聞見出し1985年
朝日新聞見出し1986年
朝日新聞見出し1987年
朝日新聞見出し1988年
朝日新聞見出し1989年
朝日新聞見出し1990年

ファイルが1年ごとになっているのは、構築当時のソフトがまだ開発中のもので、データに大量になると検索速度が低下したためである。その後、この点は改良され、ハードディスクが追加できたときは10年分を一つのファイルにする予定である。

収容データ件数は

	朝刊	夕刊	合計
81年	52,138	20,010	72,208
82年	53,789	19,767	73,556
83年	53,296	19,969	73,265
84年	55,033	21,146	76,179

1年あたりのファイルの大きさは約8MBである。

検索する年のファイルを選び、リターンキーを押すと検索画面になる。

<条件入力>

文書名：朝日新聞見出し1981年

検索条件を入力して下さい

[見出し?]
[項目1?]
[項目2?]
[項目3?]
[項目4?]
[項目5?]
[項目6?]
[全項目?]

見出し、あるいは項目別に限定したフィールド検索と、見出し、全項目を検索する全文書検索の両方が可能である。

AND、OR検索、あるいは正規表現を使つての中間一致、前方一致検索も可能である。

縮刷版見出しデータベースは、長いレンジにわたる世の中の動きを追う事などに適していると思われる。最近、深刻化している「エイズ」問題を調べてみよう。

<条件入力>

文書名：朝日新聞見出し1981年

検索条件を入力して下さい

[見出し?]
[項目1?]
[項目2?]
[項目3?]
[項目4?]
[項目5?]
[項目6?]
[全項目?] エイズ

<一覧表示>

照合条件 \$0/エイズ/
検索結果 0件

というように結果がかえってくる。

「エイズ」での検索結果は

81年 0件

82年 0件

83年 10件

< 一覧表示 >

キーワード \$0/エイズ/
検索結果 10件

1. AIDS (エイズ) __奇病AIDS
830713 朝刊 4 472 8
2. AIDS (エイズ) __ことば
830713 朝刊 4 472 8
3. 「エイズ」の話題も盛り込んで第五回国際免疫学会議__みんなの科学
830815 夕刊 5 489 1
4. 二人、改めて「シロ」 厚生省エイズ研究班__奇病AIDS問題
830820 朝刊 22 666 3
5. エイズも議題 国際免疫学会議きょうから開幕__国際免疫学会議
830821 朝刊 3 683 1
6. エイズ衰えず二千人超す__奇病AIDS
830909 夕刊 18 374 6
7. エイズへの恐れと同性愛への嫌悪感__土曜の手帳
830903 夕刊 5 125 1
8. エイズの女性患者死ぬ__保健・衛生
831009 朝刊 22 346 6
9. エイズの原因カビの一種か 国立研究所で新説__保健・衛生
831027 夕刊 18 1074 8
10. 患者は三千人超す エイズ研究会が閉幕__奇病AIDS
831126 夕刊 14 1000 8

すなわち、朝日新聞の見出しに「エイズ」という言葉が初めて登場したのは
1983年7月13日の朝刊4面だったことがわかる。

84年はやや件数が減って6件。

< 一覧表示 >

キーワード \$0/エイズ/
検索結果 6件

1. エイズの犯人判明?__保健・衛生
840424 夕刊 14 908 7
2. されどエイズ怖し__話題
841026 朝刊 7 1009 8
3. エイズの温床 公衆浴場などシスコで閉鎖__保健・衛生
841011 朝刊 22 402 7

4. エイズ治療、暗礁に 双子の血液・骨髄移植失敗__保健・衛生
841025 夕刊 18 1000 1
5. エイズ患者の輸血で? 乳児3人死ぬ 豪の病院__オーストラリア
841116 朝刊 3 595 9
6. エイズ 輸血で感染 豪は騒然__海外・喜怒哀楽
841207 夕刊 5 271 1

85年以降を参考のため、朝日新聞記事データベースのタイトル検索（見出し部分だけの検索）でひいてみると

85年77件、86年73件、87年384件、88年200件、89年177件、90年174件という結果であった。

次に「データベース」という言葉で検索してみた。

81年 1件

キーワード \$0~/データベース/ < 一覧表示 >
検索結果 1件

1. データベース化作業__集積回路
810921 夕刊 10 848 6

82年 1件

キーワード \$0~/データベース/ < 一覧表示 >
検索結果 1件

1. データベースのネット化へ実験 郵政省__データ通信
821001 朝刊 8 8 7

83年 2件

キーワード \$0~/データベース/ < 一覧表示 >
検索結果 2件

1. データベース台帳総覧作成(通産省) __情報処理産業
830330 朝刊 9 1157 10

2. データベース民間に公開を 懇談会、省庁に要望書 情報処理産業
830605 朝刊 9 159 9

84年 8件

< 一覧表示 >

キーワード \$0/データベース/
検索結果 8件

1. 「データベース」整備 振興センター発足 通産省許可 まず二十社が参加
840417 朝刊 1 599 1
2. データベース 情報化社会新語ABC
840528 朝刊 9 1007 7
3. VANやデータベース 税の減免を求める 郵政省方針 税制
840730 朝刊 3 1135 1
4. 妊娠末期の死産・生後間もない死亡 データベース化図る みんなの健康
840805 朝刊 27 185 1
5. 高度情報化社会へ報告書データベース整備急げ 高度情報化社会へ懇談会が報
841027 朝刊 1 1051 1
6. データベース・サービス業技術・資金面で援助 通産省が育成策 産業
841120 朝刊 9 753 5
7. 共同データベース提言 利用業界主導めざす 研究会報告書 情報処理産業
841218 朝刊 9 677 4
8. データベースに著作権 法整備し保護を 体系づけなど創作性 著作権
841216 朝刊 1 601 1

同様に朝日新聞記事データベースを使った検索で85以降は、85年20件、86年15件、87年25件、88年19件、89年37件、90年22件であった。

わが国における「データベース」元年をいつにするかといえ、この件数の急激に増えた1984年あたりが「元年」となるのだろうか。

次に84年におけるコンピューターの動向を探ってみた。

キーワードは「コンピューター」と「電算機」にした。

入力式で「コンピュータ」としているのは、文字列一致検索の場合、「コンピュータ」と音引きがない記述がされていると「コンピューター」では欠落してしまうからである。逆に「コンピュータ」なら「コンピュータ」でも「コンピューター」でも検索対象とすることができる。84年における検索結果は238件であった。

< 一覧表示 >

キーワード \$0~/コンピュータ|電算機/
検索結果 238件

1. 日立、互換路線変えず 社長表明 継続性を最重視 ソフト開発一時は負担
840120 朝刊 9 677 1
2. スパイ事件は「個人的暴走」__コンピューター (電算機)
840120 朝刊 9 677 2
3. 米製パソコン、日本で激突 IBM追いアップルも新型
840125 朝刊 8 856 1
4. ソフト保護法に疑義 米大使 通産相に公式表明__ソフト保護と著作権
840113 朝刊 9 441 5
5. 電算機ソフト 著作権法改正で保護 文化庁に中間報告 審議会小委
840120 朝刊 3 671 1
6. ソフト保護、やはり専門法通産省が猛反論 文化庁との対立鮮明に
840120 朝刊 8 676 1
7. ソフト保護万全を 電機業界首脳 通産省に要請__ソフト保護と著作権
840120 朝刊 8 676 4
8. 新電報システム電算機 IBMから調達 電電の国際資材調達問題
840121 朝刊 9 717 7
9. 自動車・電算機が増勢 米商務省の84年産業予測 対日競争強く意識__経済
840104 朝刊 9 109 1
10. 合併や買収大型化傾向 米電算機サービス業界__経済
840127 朝刊 8 932 8
11. インド向け電算機販売契約を破棄 ソ連懸念し米企業__貿易
840130 朝刊 3 1035 10
12. 「蒙古秘史」の言語電算機で解析研究__海外文化
840109 夕刊 5 315 9
13. (13) 電算機診断__情報多重列島
840122 朝刊 4 746 1
14. 三宅島の溶岩流 電算機で再現 京大防災研 避難情報システムに道__火山
840207 朝刊 3 217 1
15. 電算機データ生かせず 兼信の隊歴も入力__自衛官が同僚に発砲 死傷4人
840229 朝刊 23 1027 5
16. 第五世代コンピューター__情報化社会新語ABC
840213 朝刊 11 433 7
17. 米での生産まだ (コンピューター機器生産)__米国
840211 朝刊 8 374 5
18. 米の電算機部品メーカーを買収 日立金属の子会社__米国
840223 朝刊 8 772 8
19. コンピューター故障 東証二部で売買停止__株式市場
840222 朝刊 8 736 3
20. 画像処理に超LSI速さ中型電算機並み 日本電気 非ノイマン型実用化
840215 朝刊 8 504 4
21. 「IBMコピー」認める 関連役員、賞与カット 日立社長
840202 朝刊 8 44 5
22. 使いやすさで日本製が先行 高速電算機で米紙警告
840214 朝刊 8 468 8
23. 第5世代電算機 研究開発英と協力 5月にもスタート 当面、情報交換
840214 朝刊 9 469 1
24. IBM超大型機 機能強化し新モデルに__コンピューター (電算機)
840225 朝刊 8 860 8
25. 情報処理量128倍に 富士通が新ソフト IBMより低価格
840201 朝刊 8 8 1
26. 百超すユーザーが導入予定__ソフト開発・ソフト権保護
840201 朝刊 8 8 1
27. 31ビット・アドレス方式__ソフト開発・ソフト権保護
840201 朝刊 8 8 6
28. ソフト保護2省庁案まとまる 著作権法か新立法か
840215 朝刊 3 499 1
29. 文化庁 現行法内、50年間 通産省 15年、裁定制盛る
840215 朝刊 3 499 1
30. ソフト保護 通産省・文化庁の主張__ソフト開発・ソフト権保護
840215 朝刊 3 499 2
31. (解説)成長産業 縄張り争いの面も__ソフト開発・ソフト権保護
840215 朝刊 3 499 6

32. 自立へ意気盛ん(野村コンピューターシステム) __ソフト開発・ソフト権保護
840216 朝刊 8 540 4
33. 総合ソフト会社設立 __ソフト開発・ソフト権保護
840221 朝刊 8 698 4
34. ソフト権保護問題 米、新立法に反対表明 著作権条約に違反
840224 朝刊 9 817 1
35. (解説) 国内論争にも石 __ソフト開発・ソフト権保護
840224 朝刊 9 817 3
36. ソフト裁定制で対立 日米先端産業作業部会 __ソフト開発・ソフト権保護
840224 朝刊 9 817 6
37. ソフト保護は著作権枠内で 米代表、改めて強調 __ソフト開発・ソフト権保護
840225 朝刊 9 861 7
38. 今秋にもシステム監査基準 コンピューター関連の事故防止 __システム監査
840208 朝刊 9 259 5
39. 通産省、研究会設置へ __システム監査
840208 朝刊 9 259 6
40. 野村コンピューターシステム VANサービス進出 __VAN (付加価値通信網)
840224 朝刊 8 816 5
41. 教育に浸透する電算機・英国 __世界はきょう
840206 朝刊 6 180 1
42. 携帯電算機にIBMが参入 __経済
840217 朝刊 8 580 8
43. コンピューターで明代の錦織り再現 __海外文化
840221 夕刊 5 719 8
44. コンピューターの権威・坂井京大教授 __著作権
840226 朝刊 23 911 1
45. 電算機で学習能力診断 つまずきの原因ビタリ 児童対象に港区開発へ __教育
840214 朝刊 21 481 1
46. 重複出題 電算機でチェックへ __国公立大共通一次試験
840204 朝刊 3 111 1
47. 首都の守りもコンピューター時代 __警視庁
840202 朝刊 23 59 1
48. 印鑑証明も電算機で 4月にオンライン化 台東区 __二十三特別区
840210 朝刊 21 347 1
49. スーパーコンピューター __情報化社会新語ABC
840312 朝刊 11 445 7
50. 事故防止へ安全基準急げ 金融機関のコンピューター 懇談会報告 __金融
840331 朝刊 1 1193 1
51. 照明のアイデン 電算部門へ アマストと資本提携
840307 朝刊 8 244 5
52. 米大手社(コンピューター)の株半分握る 富士通
840308 朝刊 8 284 8
53. 電算機用テープにカートリッジ時代 日本IBMが新装置発表
840327 朝刊 8 1024 4
54. 超大型電算機で提携 日電と仏社 対IBM戦略 __コンピューター(電算機)
840328 朝刊 8 1066 1
55. ソフト保護 産業界でも意見分かれる 通産省と文化庁の対立反映
840303 朝刊 9 93 5
56. 伊藤忠がソフト会社 __ソフト開発・ソフト権保護
840306 朝刊 8 202 3
57. ソフト保護文化庁試案 反対の意見書提出 電子工業協会など三団体
840307 朝刊 8 244 1
58. ソフト保護「法律の形態にこだわらぬ」 米、プログラム権法に軟化?
840310 朝刊 9 373 5
59. ソフト問題、具体的に触れず 日米意見交換 __ソフト開発・ソフト権保護
840311 朝刊 9 411 10
60. “プログラム権法案”でも米が「反対」の文書 政府・自民関係者に配る
840314 朝刊 1 515 5
61. (解説) 強引な介入 内政干渉も __ソフト開発・ソフト権保護
840314 朝刊 1 515 10
62. ソフト開発の委託先を募集 __ソフト開発・ソフト権保護
840331 朝刊 9 1201 8
63. ソフトウェア 相互利用へ体制づくり きょう連絡協発足 __ソフト相互利用
840328 朝刊 9 1067 7
64. 18億円分の電算機 日立、中国から受注 __日中貿易
840316 朝刊 8 606 5

65. A T T、電算機に参入 乱戦に拍車 日本への影響大_経済
840329 朝刊 8 1106 4
66. 米への輸入禁止 I T Cが勧告へ 韓国・台湾の電算機_貿易
840301 朝刊 8 8 8
67. ソフト著作権はく奪も検討(米電算機・事務機器製造業協会副会長)_著作権
840321 夕刊 2 796 8
68. 第5世代コンピューターってなに?_みんなの科学
840305 夕刊 7 185 1
69. 雪解け水量つかむ 電算機を使って再現 洪水対策などに活用_みんなの科学
840312 夕刊 7 465 1
70. 電算機利用のシンボ_集積回路
840312 夕刊 7 465 7
71. 雨読者との対話を誘うー反論許すエッセイ 人間の存在理由は何かーコンピュ
840330 夕刊 5 1177 1
72. 日本の歌すべてコンピューターに「近代音楽館」をめざし楽譜集め_音楽
840306 朝刊 23 217 1
73. 128億円不正融資 元参事ら3人逮捕 電算機を悪用 広島農協_犯罪
840313 夕刊 11 513 3
74. コンピューター悪用有罪 懲役二年六月 猶予つき 札幌地裁判決_刑事裁判
840327 夕刊 13 1057 1
75. 第21回朝日科学教室「新世代コンピューターとロボット」_事業
840317 朝刊 22 668 8
76. 貸しビデオ店チェーン展開 電算機ネットワークで結ぶ バンダイ_ビデオ
840428 朝刊 8 1050 5
77. 米社と提携し電算機に進出 旭光学が経営多角化_コンピューター(電算機)
840404 朝刊 8 118 8
78. ソニー アップル社と提携 電算機周辺機器開発へ
840407 朝刊 8 238 4
79. より低価格のプロセッサー 日本IBMが六月発売
840414 朝刊 8 498 8
80. ソフト保護で主張すれ違い 日米先端技術作業部会_ソフト保護権
840415 朝刊 9 535 9
81. ソフト法 今国会の提案見送り 通産・文部の調整つかず_ソフト保護権
840420 朝刊 2 716 6
82. 米社通じ販路を拡大(スター精密)_パソコン
840418 朝刊 8 642 3
83. 日本電気が技術供与(YHP社に)_パソコン
840419 朝刊 8 682 2
84. コンピューターと通信回線を“結婚”させちゃう_VAN(付加価値通信網)
840414 朝刊 4 494 1
85. VAN事業部を新設(野村コンピューターシステム社)
840425 朝刊 8 918 1
86. 電算機など13億円 電電公社、A T Tと契約 対米調達、目標突破へ
840403 朝刊 3 73 1
87. 35歳限界説裏付け 電算機ソフトウェア業界 初の労働実態調査_労働
840421 朝刊 21 775 5
88. 中国・新産業革命へ電算機技術の特訓_経済
840408 朝刊 3 269 10
89. 電算機、中国と合併 エレクトロニクスで初 米パックカード社_経済
840421 朝刊 8 762 1
90. 米最新電算機「VAX」ソ連が部品入手? 米高官証言_貿易
840403 夕刊 2 100 3
91. 電算機の英語予備校_予備校
840428 朝刊 8 1050 1
92. (1)音と光の共演_ヨーロッパのコンピューター楽師
840424 夕刊 5 899 6
93. (2)市民と作曲家_ヨーロッパのコンピューター楽師
840425 夕刊 5 939 6
94. (3)音の工場_ヨーロッパのコンピューター楽師
840426 夕刊 5 979 6
95. (4)伝統と風土_ヨーロッパのコンピューター楽師
840427 夕刊 7 1029 6
96. コンピューターを使った「チケットぴあ」サービスを始めた 矢内広さん
840421 朝刊 3 757 5
97. 紙使わず電算機表示 国立リハビリセンター 盲人用に開発_身障者
840419 夕刊 14 712 1

98. ワープロ喫茶 お茶とコンピューター__にゅうす・らうんじ
840424 夕刊 3 897 1
99. 日立も空前の増益 電算機や家電好調 IBM事件響かず__3月期決算
840525 朝刊 8 886 1
100. 主要電機メーカー ニケタ増収増益 電算機・半導体 おう盛な需要
840526 朝刊 8 934 3
101. 新化学物質研究開発 電算機で効率化 通産がシステム計画__産業
840524 朝刊 9 847 1
102. 画像処理コンピューター__新製品
840529 朝刊 8 1046 10
103. ビデオゲーム機会社買収(コンピューターサービス社)__ビデオ
840522 朝刊 8 762 9
104. 国民の個人情報 国が握っている 8億8000万件 進む電算化
840504 朝刊 3 107 1
105. プライバシー保護対策、急務に__コンピューター(電算機)
840504 朝刊 3 107 4
106. 電算機要員が不足 中堅企業の6割訴え 実態調査
840506 朝刊 9 177 4
107. 第五世代電算機のまた一步 大量のデータの蓄積・検索を高速化
840510 朝刊 9 303 5
108. I C O Tが試作に成功__コンピューター(電算機)
840510 朝刊 9 303 5
109. おしゃべりパソコン(三洋電機開発)__コンピューター(電算機)
840522 朝刊 8 762 8
110. 米へ富士通の印字機__コンピューター(電算機)
840524 朝刊 8 846 10
111. 住宅地図をデータ化__コンピューター(電算機)
840530 朝刊 8 1090 10
112. 自動翻訳機売り出す 来月まず「和英」1時間に3000語__自動翻訳機
840518 朝刊 1 601 1
113. 人手修正まだ必要__自動翻訳機
840518 朝刊 1 601 1
114. (解説)異質の文法を克服__自動翻訳機
840518 朝刊 1 601 7
115. 稚アユよ!のぼれ コンピューターで漁道の流速一定に 多摩川__水産業
840515 夕刊 13 521 5
116. 大辞典を電算機化__話題
840517 朝刊 7 567 9
117. 超高速電算機 IBM会長が「実験」を確認__経済
840501 夕刊 2 26 6
118. 米ハイテク企業連合16社参加し本格活動 日本の第五世代電算機に対抗
840508 朝刊 8 224 1
119. 近づく自動検針時代 コンピューターで集中制御__くらしの科学
840517 夕刊 4 588 1
120. 超高速電算機を連結__原子力
840523 朝刊 3 801 2
121. スーパーコンピューター__原子力
840523 朝刊 3 801 2
122. コンピューター自動翻訳機を商品化した 山本武彦さん__ひと
840525 朝刊 3 881 4
123. このコンピューター時代にそろばん日本一になった 前野文之さん__ひと
840531 朝刊 3 1125 5
124. 官・民で進むオンライン化 電算機時代に警鐘 結局は人事管理の問題
840520 朝刊 23 705 6
125. 図書館 コンピューター化の周辺__家庭面
840531 朝刊 13 1135 1
126. コピー戦争 電算機もヨソクフノウ__にゅうす・らうんじ
840512 夕刊 3 401 1
127. 大地に足をつけ(村田・デジタルコンピューター社会長)__ベンチャー企業
840609 朝刊 8 320 1
128. 電算機を駆使 民間にも情報 エネルギー
840617 朝刊 3 605 2
129. 電算機間通信の国際標準化 富士通が参加の方針__コンピューター(電算機)
840605 朝刊 9 155 8
130. 将来は大市場に(日本DEC社長強調)__コンピューター(電算機)
840619 朝刊 8 682 1

ここで1番の見出しでは、字面にコンピューターという言葉がないのに検索対象になっている。

詳細画面を示せば図 5-1 の通りであり、上位項目の見出しが検索対象になっていることがわかる。

1. 日立、互換路線変えず 社長表明 継続性を最重視 ソフト開発一時は負担
840120 朝刊 9 677 1
I D……8401-1262

項目1…経済

項目2…産業

項目3…電子・電機器

項目4…コンピューター（電算機）

項目5…<

項目6…<

見出し…日立、互換路線変えず 社長表明 継続性を最重視 ソフト開発一時は負担

図5-1 詳細画面の例

6. 今後の課題

縮刷版見出しデータベース第1期計画分（1981年－1990年）は、テキスト入力も終わり、全体像が姿をみせ始めている。戦後編完成に向け、引き続き年代をさかのぼっての入力も進行中である。

非公式ながらディストリビューターへの配信打診の結果、数社からは前向きの感触を得ている。

さらに、こういったデータベースが広く世の中に広まるためには、安価で使いやすいものが必要であろうと考え、10年分をひとまとめにした電子ブック（8センチCD-ROM）化の企画も進行中である。

バブル経済崩壊後の経済不況と、当初思いもしなかった状況にまきこまれてしまったが、なんとか短期間のうちに戦後編を完成させたいと考えている。

—— 禁無断転載 ——

平成 5 年 3 月発行

発行 財団法人 データベース振興センター
東京都港区浜松町二丁目4番1号
世界貿易センタービル7階
TEL 03-3459-8581

委託先 株式会社 朝日新聞社
東京都中央区築地五丁目3番2号
TEL 03-3545-0131

印刷所 朝日新聞社印刷局印刷部
東京都中央区築地五丁目3番2号

