

データベース構築促進及び技術開発に関する報告書
十勝毎日新聞記事データベース構築

平成11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 ケー シーズ

KEIRIN

00

この事業は、競輪の補助金を受けて実施したものである。

序

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3／1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協力を促進し、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と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 東海大学教授 上條史彦氏）を設置している。

この「十勝毎日新聞記事データベース構築」は、平成10年度のデータベース構築促進および技術開発促進事業として実施した課題の一つで、当財団が株式会社ケーシーズに対して委託実施したものである。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10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で実施した課題は次表のおとりである。

平成11年3月

財団法人 データベース振興センター

平成10年度 データベース構築・技術開発促進事業委託課題一覧

分 野	課 題 名		委 託 先
社 会	1	生命保険商品情報データベースプロトタイプ構築	(株) 平和情報センター
	2	ビジネス法定文書のサンプルフォームデータベースの構築	ジャスネット コミュニケーションズ (株)
	3	グループウェアによるISO9001品質文書管理データベースツールの構築	日本電子開発 (株)
地域活性化	4	地域対応型の農業技術情報データベース構築	アスシトマイクロ (株)
技 術	5	登録ジオコードに基づく世界データベースシステムの開発	アジア航測 (株)
	6	複数の認証局間における認証データベースの有効利用に関する調査研究	(株) イフ・アドバタイジング
地域振興	7	十勝毎日新聞記事データベース構築	(株) ケーシーズ
	8	付加価値データベースシステムの開発に関する調査研究	(株) インタークラフト
	9	新規産業創出のための業種を融合した企業情報のデータベースの構築	(財) 相模原市産業振興財団
	10	地域ホームページに係わるリンク集整備および自動巡回システムと全文検索エンジンを利用した効果的検索機能の開発	ミネルバ (株)
	11	古墳・遺跡の指標となる土器データ検索パイロットシステムの作成	凸版印刷 (株)
	12	出雲古代遺跡デジタルデータベース構築	(株) 出雲王国
	13	インターネットを利用した四国地域の研究者データベースの構築	(財) 四国産業・技術振興センター
	14	長崎県観光写真素材データベースの構築	長崎メディアミックス 協同組合
	15	沖縄伝統空手・古武道国際人材リソースデータベース	(株) アイエムアイ コーポレーション

目 次

1. 概要	1
1.1 目的	1
1.2 実施内容	2
1.2.1 データベース構築	2
1.2.2 ハードウェア及びソフトウェア	5
1.2.3 作成	5
1.3 実施体制	7
1.3.1 実施経過	7
2. 実施と結果	8
2.1 実施の詳細	8
2.1.1 マイクロフィルム化	8
2.1.2 デジタル画像化	8
2.1.3 デジタルテキストデータ化	10
2.1.4 インデックス作成	12
2.1.5 全文検索	13
2.1.6 リサーチ	14
2.1.7 データベースと環境	28
2.2 問題点	29
2.2.1 製作各段階での問題点	29
2.2.2 全体の問題点	31
3. 今後の課題	32
3.1 膨大なデータ量	32
3.2 当事業を通じて、問題の所在が明らかになった	32
3.2.1 証明できたこと	32
3.2.2 その他の問題	34
3.3 新たな可能性も見えてきた	34
3.3.1 可能性の一つはインターネット接続	34
3.3.2 もう一つの可能性	35
3.3.3 終わりにあたって	35

図 表 目 次

図 1 - 1 システム工程図	6
-----------------	---

十勝毎日新聞記事データベース構築

1 概要

1.1 目的

十勝地域の、最大情報源である十勝毎日新聞の過去の記事を、使いやすいデータベースに構築することを、目的とする。

最終目標は、十勝毎日新聞が、80年にわたって蓄積してきたすべての新聞記事をデータベース化することであるが、それに先立ち、当事業では、十勝毎日新聞記事を一定期間に限って、地域政治・産業・流通・暮らし・文化交流等を、円滑にするための、データベース構築をすることを、目的とする。さらに、この事業を通して、データベース構築の、さまざまな問題点を明らかにし、その解決策と、今後の課題をも、検討することをも、あわせて目的とする。

当事業は、北海道・十勝地域で、高度情報化社会に不可欠なデータベース構築を、行った事業である。データベースは、十勝毎日新聞記事で、構築した。十勝毎日新聞は、帯広に本社を置く、80年の歴史を有する、地域新聞である。十勝地域では、部数8万3000部の、部数1位の主読紙でもある。各家庭に配布される新聞は、産業や経済、地域政治から、暮らしのすみずみまで、役立つ日々の情報源として、広く活用されている。地域に暮らす人々が、情報源として、十勝毎日新聞紙面に寄せる、期待と信頼は大きい。

しかし、活用されているのは、日々の新聞紙面であり、過去の記事ともなると、新聞縮刷版もしくは、もしくは図書館等の保存新聞原紙、あるいは新聞社に備え付けられたマイクロフィルム等を利用する以外、検索する方法のないのが現状であった。十勝毎日新聞の過去の新聞記事を、誰にでも使える、データベースとして構築することができれば、さらに有用な地域の情報源として、さらには、歴史的な情報源としても、価値を持つようになると、予想された。ただし、新聞記事のデジタルデータベース構築は、最新のコンピュータ機器と、専用ソフトを開発使用しても、膨大な作業時間とコスト高が、容易に予想できた。そのため、新聞記事のデータベース構築の試みは、本格的に、取り組まれることが無かった。

高度情報化社会にデータベース構築は不可欠であり、十勝では、十勝毎日新聞の記事がデータベースの情報源としてふさわしいと判断し、今回、これを使って、期間を限定し、新聞記事のデータベース構築事業に、取り組むことにした。市販のコンピュータ機器、市販のソフトウェアを用いて、新聞記事のデータベース構築に、障害となる問題の所在を明らかにすることを、目的とした。

その事業の、目的、実施内容、実施と問題点、今後の課題を、ここに報告する。

1.2 実施内容

1.2.1 データベース構築

当事業では、1997年1月から12月までの十勝毎日新聞紙面を用いて、データベースのデジタル画像化を行った。地域新聞のデータベースを構築し、全文検索を可能にさせるために、次の工程が必要であり、当事業では、それを、実施した。

(1) マイクロフィルム化

データベース構築に必要な、所定の新聞紙面を撮影してマイクロフィルムを作成した。十勝毎日新聞社では、新聞紙面を定期的にマイクロフィルム化しているので、当事業では、1997年1月から12月分のマイクロフィルムを、使用した。

(2) デジタル画像化

マイクロフィルムをT I F F化して、デジタル画像化した。データベースの性能比較のため、1月から3月までを400dpiで、2月から12月までを200dpiで、T I F F化した。

(3) デジタルテキストデータ化

デジタル画像化された、新聞紙面の文章部分を、O C Rを利用して、デジタルテキストデータ化した。

(4) インデックス作成

作成したデジタルテキストを元に、全文検索用のインデックスを作成した。

(5) 全文検索化

データベース構築された所定の新聞記事（1997年1月～3月）の、全文検索を実現した。

(6) 試用

完成したデータベース構築の全文検索を試用した。

(7) リサーチ

一般人の試用に供しアンケートをとった。

実施内容の詳細は以下の通り

(1) マイクロフィルム化

1997年1月から12月までの、十勝毎日新聞紙面のマイクロフィルムを、使用した。

(2) デジタル画像化

マイクロフィルムのアナログ画像を、スキャニングし、T I F Fデータ化して、この段階で、新聞紙面の、デジタル画像化をした。当事業の完成状況を予測して、比較検討出来るように、データ容量と作業スピードを考慮し、新聞紙面の1997年の1月から3月分を、400dpiでT I F F化した。2月から12月分を、200dpiでT I F F化を、行った。T I F F化作業を終えた段階で、性能比較を行った結果、400dpiがふさわしいと判断し、これを用いて、データベースを構築することに、決定した。1月から3月までの、3ヶ月間の400dpi画像で、全文検索を行えるようにすると、決定した。

当事業では、新聞紙面をT I F Fデータ化するための、一連のシステムを、作成した。なぜなら、新聞記事の分量は膨大な量が控えており、データベース構築の処理スピードを上げることが、重要であると判断したからである。当初の予測では、T I F F化してデジタル画像を作成する、この作業がいちばん時間がかかると思われた。そのために、この作業を短縮化するための一連システムを、他の作業に優先して作成した。しかし、結果的には、他の作業で予想外の時間がかかった。

(3) デジタルテキストデータ化

デジタル画像化した新聞紙面の文章部分を、OCRを利用して、デジタルテキストデータ化した。このデジタルテキストデータは、全文検索のインデックス(索引語)の元になるデータである。

T I F F化するためのスキャナーは、米サンライズ社の「S R I P 5 0」を使用し、作成したT I F Fデータからテキストデータを生成するためのソフトには、認識、精度等、後述の全文検索エンジンとの連携等を考慮し、市販の「パナOCR」を採用した。ただし、新聞という極めて複雑なレイアウトを持つメディアを効率良く処理するためには、若干の機能不足があり、それらを補完するツールプログラムを作成し、使用した。この工夫により、大量のT I F Fデータからテキストデータを自動的に作成することが可能になり、生産効率が向上した。

(4) インデックス作成

作成したデジタルテキストを元に、全文検索用のインデックスを作成した。この作業段階では、認識精度、高速性などから判断して、市販の全文検索エンジン「パナリサーチ」を使用した。

(5) 全文検索化

データベース構築された新聞記事の、全文検索を実現した。1997年1月から3月までの400dpiにデジタル画像化された、十勝毎日新聞記事を元に、全文検索可能なデータベース構築を完成させた。

(6) 試用

完成したデータベースの試用を行った。新聞記事データベースの全文検索の、性能確認、課題等を検討した。

(7) リサーチ

完成したデータベースを一般人への試用に供し、リサーチを行って、ユーザー評価を得た。

1.2.2 ハードウェア及びソフトウェア

上記、実施作業で使用したハードウェア及びソフトウェアは、以下の通り。

(1) ハードウェアおよびOS

(A) T I F Fデータ作成 「S R I P 5 0 (P e n t i u m 1 5 0 M H z、M E M 6 4 M B、O S W i n N T 3 5 1 英語版)」

(B) 作業用 「P e n t i u m 1 5 0 M H z M E M 3 2 M H z」×1、「P e n t i u m 2 0 0 M H z M E M 4 8 M B」×2、「P e n t i u m 3 0 0 M H z M E M 9 6 M B」×1

(c) 開発用 「P e n t i u m 1 6 6 M H z M E M 4 8 M B」×1、「P e n t i u m P r o 2 0 0 M H z M E M 4 8 M B」×1

(D) C D - R 作成 「P e n t i u m 9 0 M H z M E M 3 2 M B」×2

※いずれもネットワーク環境

(2) ソフトウェア

(A) 画像処理ソフト系 「V i d a r」、 「S c a n f i x」

(B) O C R ソフト 「パナO C R」

(C) 全文検索エンジン 「パナサーチ」

(D) グループウェア 「パナピオス」

(E) グラフィックビューアー 「パナビューアー」

(F) インターネットブラウザ 「ネットスケープナビゲータ」

(G) リレーショナルデータベース 「オラクル」

(H) テキストスクリプト言語 「A W K」

(I) 開発言語 「V i s u a l B a s i c D e l p h i」

1.2.3 作成

新しく作成したプログラムおよびソフトは、以下の通り。

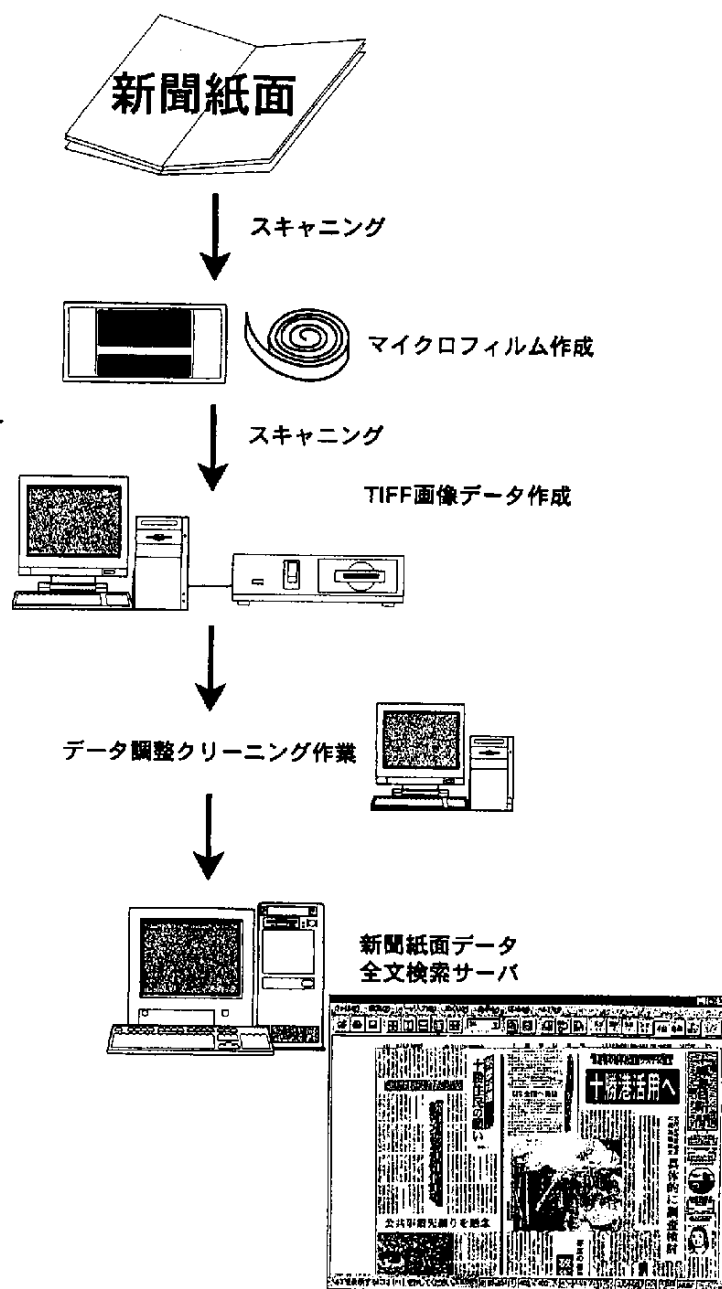
(1) 新聞専用画像処理ソフト「リネマー」

(2) ファイル名チェックスクリプト「D a t e C h e k e r」

(3) ページ数チェックスクリプト「P a g e C h e k e r」

(4) O C R一括処理プログラム「O C R B a t c h」

(5) T I F F・テキスト一括登録プログラム



■十勝毎日新聞記事でデータベース構築を行った。

■1997年1月から12月までの既存マイクロフィルムを使用した。

■スキャンニングの自動化ソフトを作成した。

■アナログ画像をTIFFデータ化してデジタル画像を作成した。

■デジタル画像の文章部分をOCRでデジタルテキストデータ化した。

■1997年1月から3月分の全文検索用のインデックスを作成した。

システム工程図

図1-1

1.3 実施体制

当事業を実施するにあたり、(株) ケーシーズでは、平成10年月に、プロジェクトチームを結成した。チームメンバーは、ディレクター1名、システム管理者1名、入力責任者1名、入力業務者1名、計4名で、構成した。

実施にあたり、システム全般に関しては、九州松下電器(株)と連携して行うこととした。また、記事提供に関しては、十勝毎日新聞社の全面的な支援を得た。

1.3.1 実施経過

- ・ 1998年7月21日 委託契約を締結する
- ・ 7月27日 プロジェクト会議開催
- ・ 8月3日 十勝毎日新聞マイクロフィルムよりスキヤニング開始
(1997年1月～12月分)
- ・ 9月18日 スキヤニング終了 デジタル画像作成開始
- ・ 10月30日 スキヤニングデータをCD-ROMに焼き付け
(1月～3月 400dpi 2月～12月 200dpi)
400dpiデータ(1月～3月分)を全文検索用に決定
- ・ 11月5日 十勝管内図書館に当事業紹介
- ・ 11月24日 OCRでテキストデータ読み込み開始
- ・ 12月4日 全文検索でインデックス作成開始
- ・ 1999年2月 データベース構築作業終了
(デジタル画像1997年1月～12月分収録)
(全文検索 1997年1月～3月分)
- ・ 2月10日 試用運転、アンケート実施
- ・ 2月15日 報告書作成

2 実施と結果

実施した内容と効果、その結果を報告する。

2.1 実施の詳細

2.1.1 マイクロフィルム化

当事業では、すでに十勝毎日新聞社で作成したマイクロフィルムを、使用した。データベース構築を前提とした、新聞紙面のマイクロフィルム化作業の、注意点を記す。

新聞紙面は、ページにより、印刷状況に差があるので、それを考慮して、撮影フィルムの濃度が一定になるように、注意する。とくに、文字部分が鮮明に表れるように撮影する。スキャニング作業とデジタル化（T I F F化）処理を前提として、撮影したフィルムが、全体的に均一な画質になるように、注意する。

※当事業では、1997年の新聞を使用した。黄ばみのあるような古い新聞を、データベース化する場合は、この段階の作業に、相当時間がかかると予想された。

2.1.2 デジタル画像化

(1) マイクロフィルムをT I F F化して、デジタル画像化した。

マイクロフィルム上のアナログ画像を、スキャニングし、T I F Fデータ化して、この段階で、デジタル画像化をした。データ容量と作業スピードを比較検討するために、新聞紙面の

1月から3月分を、400dpiでT I F F化した。2月から12月分を、200dpiでT I F F化を、行った。

(2) 当事業では、新聞紙面をT I F Fデータ化するための、一連のシステムを、作成した。

作成したシステムの内容は、以下の通り。

- (A) ファイル名変更プログラム
- (B) 画像編集ツール各種
- (C) 上記2つは、後に「リネーマー」として統合した

T I F F化のためのシステムを作成したのは、新聞記事の分量は膨大な量が控えており、データベース構築の、処理スピードを上げることが、重要であると判断したからである。

当初の予測では、この作業が、いちばん時間がかかると判断した。そのために、この作業を短縮化するための一連システムを、他の作業に優先して作成した。

(3) 画像ノイズの除去には注意を要した。

デジタル画像化の作業段階で、後に控えている文字読みとり作業の、OCRの確率を高めるために、画像のノイズ（ゴミ）を、目視で除去した。

必要以上にノイズ除去すると、ひらかなの濁点や漢字の点までも除去してしまったり、写真が汚くなるので、全文検索時のOCRの確率を高めつつ、かつ、イメージ画像を閲覧する時の高品位を保つべく、マンパワーの目視による、スキニングの調整が必要であった。

(4) デジタル画像の画質には、細心の注意を払った。

なぜなら、新聞記事のデータベースの完成後、閲覧使用する際の検索記事のヒット率の高さ、操作の容易さなどは、すべてデジタル画像の品質にかかってくると、判断したためである。

予測どおり、この作業が、マンパワー作業では、もっとも時間を要する作業であった。その理由は、次の要因に拠る。

TIFF化処理されたデータ形式は、デジタルであるが、最終的には、全文検索で閲覧する人には、見え方としてはアナログである。データベースの、快適な使用状況を考えると、鮮明な画像が求められる。構築されたデータベースを閲覧する際、鮮明な画像に見えるように、デジタル画像作成のもっとも初期のTIFF化の作業段階で、視認で、画像の良し悪しを、判断しなければならなかった。

(5) TIFF化は、画質の良し悪しを決定する段階であった。

しかも、人間の視認判断を介在させなければならない作業で、これが故に、データベース構築の、全工程の完全自動化はできなかった。

※マンパワーの長時間作業というのが、おおきな問題点であった。

将来的には、人間の視認に近い判断能力を備えてソフトが開発されれば、アナログ・テキストのデータベース化は、完全自動化が実現する。

(6) スキニングの自動化ソフトを作成した。

検索して表示されたデジタル画像の新聞紙面が、モニターにゆがんで表示されると、閲覧者に視覚的なストレスの発生が予想されるので、スキニングの段階から、注意を要した。

スキニングの際の、正確な画像の中央寄せ・トリミング・余白部分のノイズ消去などを施した。この段階の作業を、当初はマンパワーで行っていたが、コンピュータ処理で、中央寄せ・トリミング・余白の除去などが、自動的に行えるソフトを、作成した。この結果、スキニングの精度が高まり、作業のスピードアップも実現した。

※比較的新しい新聞を撮影したマイクロフィルムの場合には、問題は少ないが、マイクロフィルムに納められた古い紙面の場合は、紙面の撮影位置が一定でないことも多く、そのため、このスキヤニング調整作業に、相当時間がかかると予想される。

(7) 時間短縮のため複数台のコンピュータを使用した。

T I F F化されたデータは、別のコンピュータで、チェック作業を行い、ノイズ（画像のゴミ）除去や中央寄せの作業を、行った。当事業では、所定の新聞紙面の、T I F F化作業（デジタル画像化作業）は、自動化のソフトを作成以後、一括作業として行った。

さらに、スキヤニング・データ処理が自動化できるようになったので、複数コンピュータ（3台）を使用して、時間短縮をはかった。

(8) 効率のよいファイル名をつけた。

スキヤニングおよびT I F F化の段階で、テキストデータ化するプロセスや、検索のプロセス、データのメンテナンス性を考慮し、日付とページ数を読み取れるファイル名に変更する作業を行った。

(9) 高品位な画像づくりに十分なマンパワーを投入した。

当事業では、新聞紙面の良質なデジタル画像化を重視したため、必要とされるマンパワーを十分に投入した。さらに、T I F F化の作業には、必要なソフトを作成し、細心の注意を払って作業を行ったため、高品位のデジタル画像を得ることができた。マイクロフィルムのアナログ画像が、T I F F化されて、高品質のデジタル・イメージ画像になって、ストックされた。

(10) ソフトを統合してシステムとした。

デジタル化作業では、部分的に効率アップをはかるソフトを開発した。さらにスピードアップを図るため、個々に作成したソフトが、全体的に効率よく機能するためのプログラムを組んだ。これによって、一連の作業ができる、T I F F化のためのスキヤニング・システムとして統合された。

2.1.3 デジタルテキストデータ化

デジタル画像化された、新聞紙面の文章部分を、O C Rを利用して、デジタルテキストデータ化した。デジタル化されたテキストは、後の作業で全文検索のインデックス（索引語）の元になるものである。

(1) 全文検索に絶対必要なのが、検索用に完備したテキストデータである。

デジタル画像化された新聞紙面に、正確に対応したテキストデータが無ければ、精度の高い全文検索は望めない。データベース構築の質を決定づけるものが、この検索用のテキストデータである。そのために、T I F F化されたデジタル画像から、テキス

トを正確に読みとるOCRが必要だった。

(2) T I F F化された画像データをもとに、検索用のテキスト・データを作った。
この過程で、高信頼性・高速と評価されている、九州松下電器が開発し市販している「パナOCR」を選んだ。認識率の高さ、スピード等、各種の各種のすぐれた機能をそなえたOCR。それを活用して、さらに、これを新聞用のOCRとして機能させるために、各種のプログラムを組んだ。

(3) OCRのプログラムの流れを記すと、以下の通りである。

(A) まず、T I F F化された、画像を開く。

(B) OCRが、新聞紙面のデジタル画像に対して、レイアウト認識をする。

OCRソフトは、新聞紙面を、いくつかの領域に自動的に分割し、レイアウト認識をする。

分割された領域を、これは文字の領域、表の領域、それからイメージの領域と判断する。さらに、イメージ領域の写真とか挿し絵、あるいは空白領域と判断する。

(C) レイアウト認識をすると、文字の領域と表の領域だけを文字変換の対象とする。

(D) 文字領域を選択した後に、縦書きか、横書きかを、認識する。

文字間隔、行間隔などから、自動的に認識判断し、それぞれの領域に、縦書き、横書き、という属性を付与し、番号をつける。

その属性、番号に従って、次のプロセスの文字認識作業を行う。

したがって、新聞紙面の読み取れる文字はすべて読みとるのではあるが、読みとられた文章は、新聞記事の文章の流れではなく、OCRの付与した番号の流れになる。

(E) 文字認識は、次のように行う。

(F) レイアウトで分類された属性にしたがって、各領域に付番された順番に、文字認識される。

(G) 今回、使用した、「パナOCR」では、横書きの領域から、先に読み始める。

それがすむと、次に、縦書きの領域を読んで行く。

さらに、この読み進む流れは、左上から右下と方向が決められていた。

このOCRソフトは、左綴じの横書き文書を想定して作成されていた。

(H) このOCRは、連続登録機能が使えないので、新聞用のOCR用に、連続登録を自動的に処理するツールや、ソフトが必要だった。

(I) 使用したOCRの連続登録機能は、次のようにして解決した。

(a) 使用したOCRでは、マンパワーで、1ページずつ、オペレータが作業処理をしなくてはならなかった。しかし、今回のデータベース構築作業では、前段のT I F Fデータ化するところで、各ページには、ユニークな重複しないファイル名を付けておいたので、そのファイル名の拡張子を、<.TXT>に変更するだけで、何月何日何ページ目のテキストデータであるか明示できるため、一括処理できるツールの規則性を確立させることができた。これにより、OCR作業には、オペレータが付いている

必要が無くなった。

(b) 一括処理できるソフトは、OCRの内部が非公開のため、原始的ではあるが、外部からOCRを制御する方法を採った。

(c) ただし、その動作においては確実性が保証されていないので、処理中のコンピュータでは他の作業が行えなかった。連続している他の作業は、独立した他のコンピュータを使用する体制をとった。

これらの作業は、TIFFデータさえ用意してあれば複数のコンピュータを利用できるため、最大6台のコンピュータを併走させた。

(4) 活用した「パナOCR」は、内部仕様が非公開のため、外部から制御するソフトを作成した。外部から制御する方法とは、あたかもキーボードを使って操作指示を出すのと同じ効果を持つコマンドを、プログラムにによって発行する方法である。

最終的には、この市販の一般用OCRに、外部から制御できるソフトを組むことで、新聞用OCRとして、一括処理できる機能を持たせることができた。

(5) 今回作成したのは外部から制御するソフトで、OCR一括処理プログラムである。

具体的に内容を記すと、次のようになる。

(A) 何月何日の分として、作成したデジタル画像と、それに対応するテキストデータが整合するよう、ファイル名に規則性を持たせた。

その登録を、一連の作業で効率よく行うための、ソフトを作成した。

(B) 通常紙面の他、号外や版の違いなどを区別するルールを作った。

そのルールに従って、自動的に処理するソフトを作成した。

(C) コンピュータが判断処理できて、そのことで人手が省略できる処理は、すべてコンピュータに任せるソフトを作った。各種のツールも開発した。

(D) 作成した各種ソフトを、システムとして統合したことにより、この段階での、人為的な間違いが減少するという副次的な効果があった。

(E) それでもなお、全作業中に3回ほど、同じ紙面があるというミスが生じた。こうした人手による発生しがちな入力ミスのチェックには、スクリプト「awk」を活用することで効果を上げた。

(6) デジタルテキストデータ化の段階でも、コンピュータの自動処理ができる、可能な限りのシステム化を図った。

2.1.4 インデックス作成

(1) 作成したデジタルテキストを元に、全文検索用のインデックス(索引語)を作成した。ここでは、認識精度、高速性などから判断して、市販の高性能全文検索エンジン「パナサーチ」を使用した。

(2) この作業段階に大きな問題があった。

予想できなかったインデックスの長時間作業である。

(A) 高速で検索できるようにするための、インデックスを作成した。

インデックスは、全文検索エンジン「パナサーチ」が、自動的に作成した。

(B) インデックスは、以下のプロセスを経て、自動的に作成されている。

(C) テキストファイルを構文解析して、理論的な整合性を持たせ、分かち書きをする。それぞれの単語にコードを付与し、インデックス（索引語）として特性を与える。

(D) この処理に、予想外の時間がかかり、データベース構築作業の、二つ目の大きなネックになった。インデックスの作成には時間がかかるので、前もって十分な期間を割いて、スケジュールに組み込んでおいたが、その予想が大きく外れた。

(E) インデックス作成は、全文検索をいかに早く行わせるかという、システムの根幹に関わる問題であって、全文検索エンジンの性能を決定づけるものである。それだけに、早く検索できる仕組みづくりには、全文検索ソフトを製作している各社は、工夫をこらして高性能化、差別化をはかっている。素早い全文検索を可能とする高度な仕組みのプログラムを組めば、それだけインデックス作成の作業時間が必要となる。

(F) 使用した全文検索エンジン「パナサーチ」も、この作業に時間をさいている。

(G) データベース構築作業の後半では、読み込んだテキストファイルを、24時間体制で、インデックス化する作業が続いた。

(H) その時間中に、処理済みの部分のテストのために、全文検索を行うと、レスポンスが遅くなった。

(I) 使用した全文検索エンジンは、少量のデータベース処理を想定しているためか、インデックス作成の自動処理の進み具合を表示するものがなかった。

膨大なインデックスを必要とする今回は、工程管理を立てにくかった、という問題も発生した。

2.1.5 全文検索

データベース構築された所定新聞記事の、全文検索を実現した。

(1) 1998年7月から、データベース構築作業を開始し、1999年2月をもって、十勝毎日新聞記事を元に、全文検索可能なデータベース構築を完成させた。

データベースの元になるデジタル画像は、1997年1月から12月まで作成した。全文検索ができるデータベースは、1997年1月1日から3月28日までである。

(2) インデックスと、全文検索の対象となるテキストデータは、カテゴリーごとに階層構造となっているところの、キャビネット（本棚）のドロア（引き出し）のフォルダ（ファイル）の中に、保存した。ただし、今回使用した全文検索エンジンでは、いちばん上位のキャビネット単位で、データ容量999MBまでという制限があった。

新聞の画像データを収納する際は、999MBの制限に納まるような階層構造として保存しなければならなかった。

1997年度の、各月の新聞のページ数は、1ヶ月分で約780ページ。

それは、400dpiの画像データで、800MB弱のデータ量である。

1キャビネットは999MBの容量があるので、新聞1ヶ月分を収納した。

データ収納が階層構造になっているため、次のような手法をとった。

1997-1というキャビネットを作り、その中に1997-1というドロアを作り、さらにその中に1997-1というフォルダーを作って、ここに1ヶ月分のデジタル画像データを収納した。当事業では、3ヶ月分のインデックスとテキストデータ、新聞の画像データを、16GBのハードディスクに、収納できた。

(3) 現時点の環境では、インデックスとテキストデータをハードディスクに収納して、新聞紙面の画像データは、CD-ROMやDVDなどの外部の大容量の記憶媒体に収納し、オートチェンジャー等を活用することが、現実的手法であると考えられる。将来の本格的運用の場合、十勝毎日新聞なら80年分の新聞紙面をデータベース化する場合、デジタル画像の収納は、新たに考える必要がある。

(4) 画像データとテキストデータが対になって保存され、インデックスが作成された時点で、全文検索は機能する。当初、まるまる3ヶ月分のインデックスを作成する予定であったが、全文検索の処理速度の見込みがつかず、予測外の時間がかかり、しかも、全文検索エンジンの仕様上、進行状況が把握できないこともあって、約1200時間(12月4日～2月10日)を要して、約3ヶ月分弱のインデックス作成にとどまった。現時点(2月現在)では、新聞記事データベース(1997年1月1日～3月28日分)の、インデックスが完了した。

(5) 記事検索を行うと、インデックスの生成処理が停止している場合には、2ヶ月分のデータに対しては、1秒足らずで、目的記事の所在を表示することが、できた。その表示にしたがって、目的紙面の画像を、やく15秒で、呼び出すことが、できた。

(6) すべての実施作業を完了し、ここに、十勝毎日新聞データベース構築を、完成させた。

2.1.6 リサーチ

完成したデータベースの試行利用を行い、性能確認、試用評価、検討課題等を、リサーチを行った。

(1) 試行利用の運用とリサーチは、2月10日から12日の3日間、十勝毎日新聞社のロビーで、行った。対象者は一般人とした。説明、デモンストレーションを行った結果、一般人より17名のアンケート回答を得た。

— 15 —

(3) アンケート結果を集約すると、次の評価が得られた。

ただし、リサーチの際に、バックグラウンドでインデックスを継続作成しているという条件下であった。そのためレスポンスに関しては、評価が低めに表れている。

(A) 十勝毎日新聞記事データベース・アンケート集計結果

(回答箇所は■印をつけた)

■記事検索について、お聞かせください。

- ・目的の記事は、見つかりましたか？ (■はい 14 名・■いいえ 3 名)
- ・見つけるのは、簡単でしたか？ (■はい 11 名・■いいえ 5 名)
- ・見つけるまでの時間は、早かったですか？ (■はい 4 名・■いいえ 10 名)
- ・どれくらい時間が、かかりましたか？ (■1～2分 5 名、■3～5分 5 名、■6分以上 1 名 45 秒 1 名)

■操作に関して、お聞かせください。

- ・操作は、簡単でしたか？ (■はい 12 名・■いいえ 5 名)
- ・むずかしかったのは、どの操作部分でしたか？
(■キーボード操作、■キーワードの付け方、■検索テキストの読みとり、■ファイル番号の判別)
- ・検索の単語は、3 つでしたが、適当でしたか？ (■はい 11 ・■いいえ 2)
- ・いくつ、使いましたか？ (■1 つ 2 名・■2 つ 7 名・■3 つ 6 名)
- ・もっと多い方がよい (4 つ・5 つ・■6 つ 2 名)

■新聞紙面の表示について、お聞かせください。

- ・紙面が表示されるまでの、時間は早かったですか？ (■はい 3 名・■いいえ 13 名)
- ・どれくらい時間が、かかりましたか？ (■1～2分 7 名、■3分 1 名、■40 秒 3 名)
- ・表示された紙面は、読みやすかったですか？ (■はい 16 名・■いいえ 1 名)
- ・紙面を、拡大してみましたか？ (■はい 13 名・■いいえ 3 名)

■プリントアウトについて、お聞かせください。

- ・プリントアウトを、してみましたか？ (■はい 9 名・■いいえ 5 名)
- ・操作方法は、簡単でしたか？ (■はい 8 名・■いいえ 1 名)
- ・どれくらい時間が、かかりましたか？ (■1～2分 4 名、■3～5分 2 名、■6分以上 1 名)
- ・プリントアウトの文字部分は、満足できましたか？ (■はい 7 名・■いいえ 1 名)
- ・プリントアウトの写真部分は、満足できましたか？ (■はい 8 名・■いいえ 1 名)

■全文検索について希望を、お聞かせください。

- ・今後とも、お使いになりたいですか？ (■はい 14 名・■いいえ 2 名)
- ・どんな目的で、お使いになりますか？

(■探し物、■調べ物、■情報収集、■お店探し、■レシピ調べ)

- ・どの部分を、強化すればいいですか？

◇操作性 (■もっと簡単に、■絞り込み検索できるように)

◇スピード (■もっと早く)

◇表示画面 (■プレビュー画面が欲しい、■はじめから紙面表示を)

◇プリントアウト (■スピードアップ)

◇ネットワーク仕様 (■インターネット検索)

◇データ量 (■圧縮送信できるように)

6 この新聞の全文検索を、改良し、発展させたいと考えています。

上記アンケートのお答え以外に、お気づきのことや、

ご希望があれば、ぜひ、教えてください。

■表示仕方に工夫を。

■検索用のヘルプが欲しい。

- 検索済みは色を変える。
 - テキストの正確な変換。
 - 役所の情報の全文検索にも使えるように。
- アンケート、ご協力ありがとうございました。

(4) アンケート集計結果の分析

アンケートに応えたモニター１７名は、新聞記事の全文検索システムを、はじめて操作する人たちであった。

(A) 記事検索のヒット率は高く８２パーセントを示した。操作そのものは簡単と答えが７０パーセントと高かったものの、パソコンの習熟度のちがいによるヒット到達率にバラつきがあった。

(B) モニターへの新聞紙面の表示に関しては、遅いと答えた人が７６パーセント。高速コンピュータの使用を求むる声が高かった。

(C) モニターに表示された新聞紙面と、そのプリントアウトには、ほぼ全員の９４パーセントが満足と答えている。検索は簡単だが、表示が遅い。ただし表示された新聞紙面とプリントアウトには満足という、明らかな、結果がでた。

(D) 現状での能力でも、今後とも実用に使いたいと望む声も８２パーセントと高かった。紙面表示のスピードアップがなされると、実用使用の要望はいっそう高まると予測される。

(E) 新聞記事をデータベースとして活用したい人は多く、当事業で開発実施した、十勝毎日新聞記事データベースの全文検索システムを、実用使用したい人からは、具体的な要望も出されていた。その代表的なものは、インターネット等での、ネットワーク上でも、使えるようにして欲しいというものであった。

(F) 当事業の成果物『十勝毎日新聞記事データベース』を、現状でも「今後とも使いたい」という８２パーセントのポイントが、当事業への評価と判断している。

(5) 当データベースの全文検索の仕様は、以下のとおり

オリジナルデータ

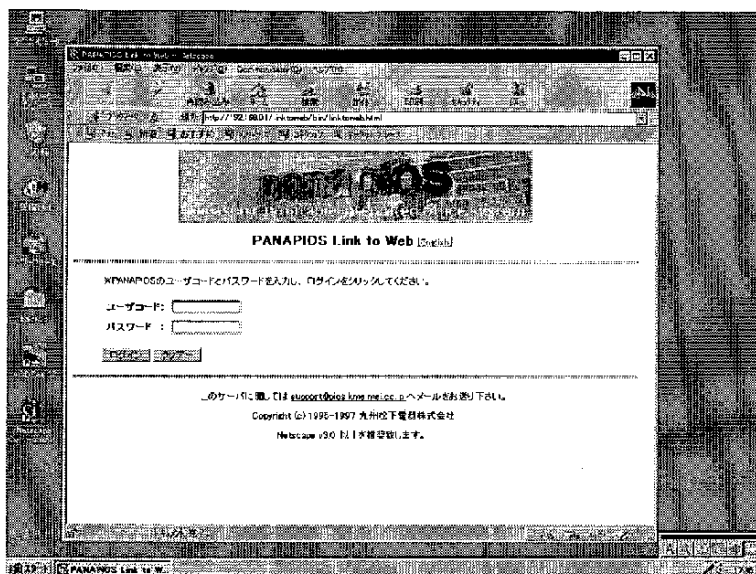
データ容量 / T I F F 約 2.4GB、テキスト 120MB

仕様ハード / P e n t i u m 2 400MHz MEM 128MB、ハードディスク 16GB、NT4.0 t S P 3

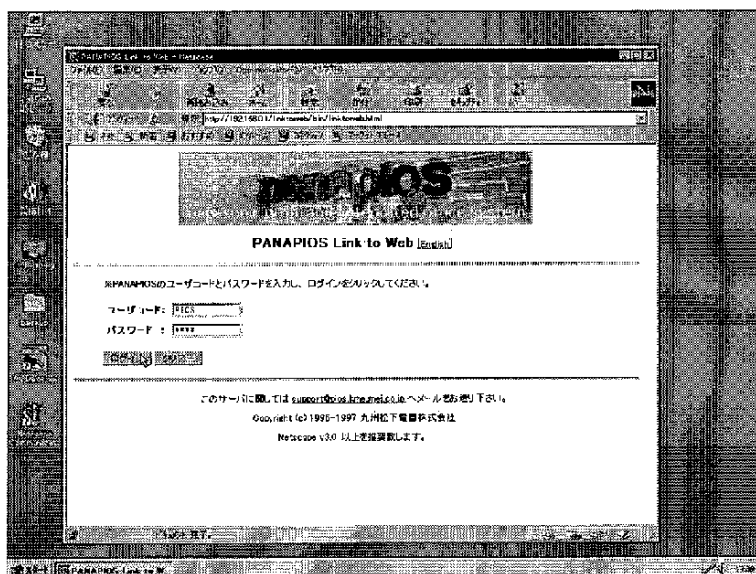
(6) 全文検索手順

データベース全文検索の、モニター画面と、使用方法は、以下のとおり
なお、探したい記事は、十勝川の環境関連記事であった。

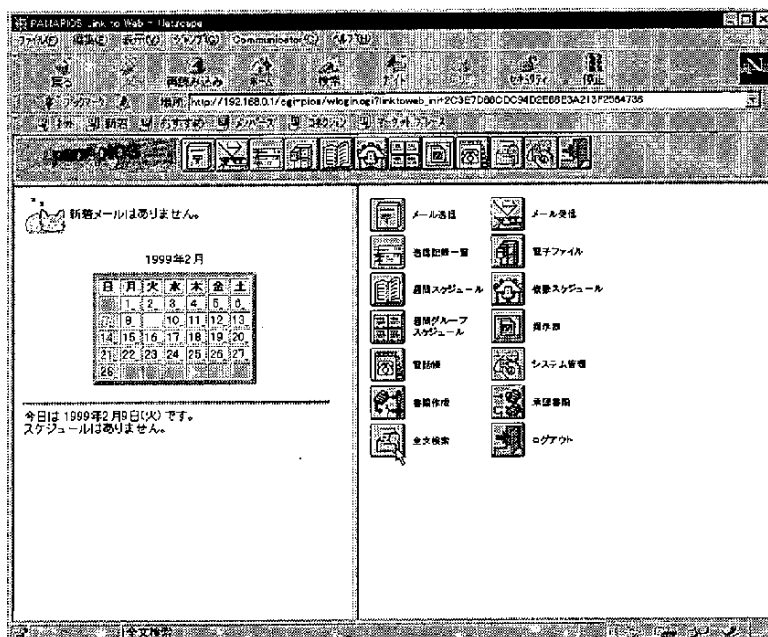
(A) インターネットのブラウザを用い、「パナピオス」のログイン画面を開く。



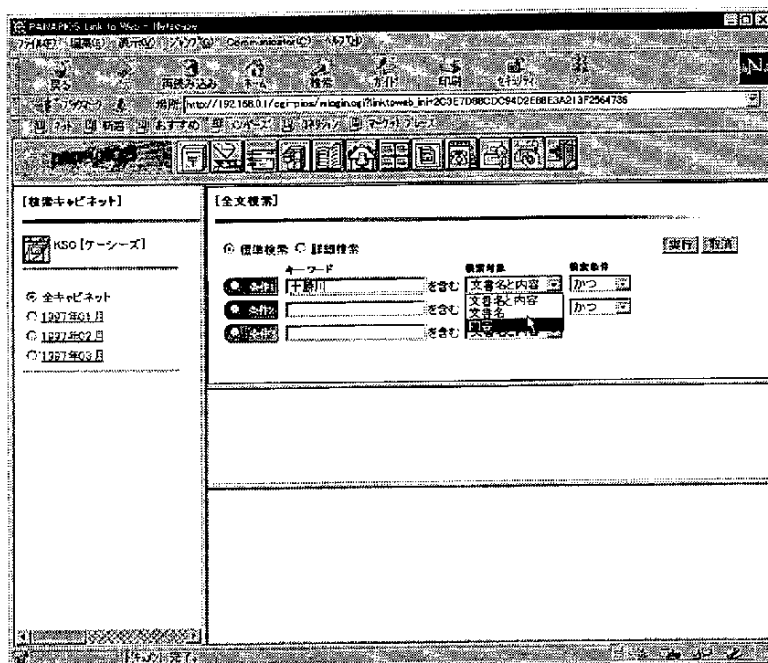
(B) ユーザーコードとパスワードを入力し、ログインをクリックする。



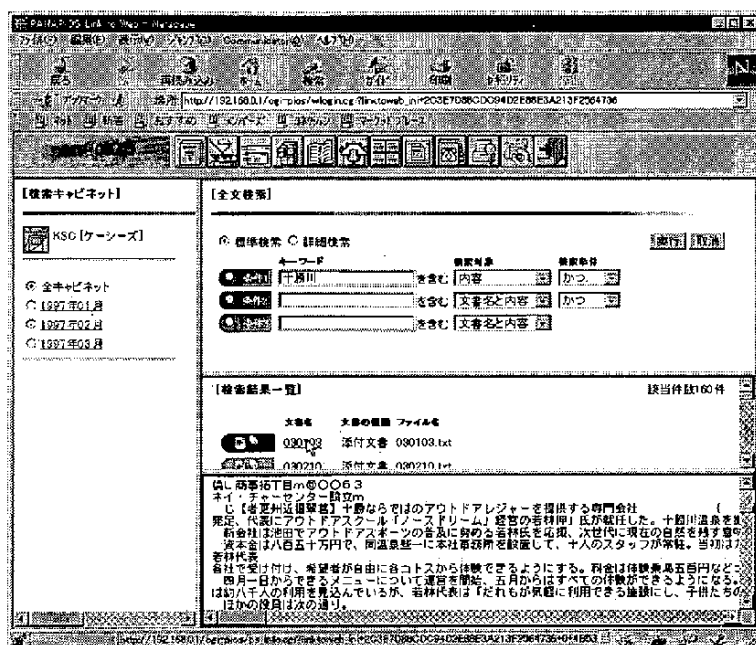
(C) メニューの中から、全文検索を選択し、クリックする。



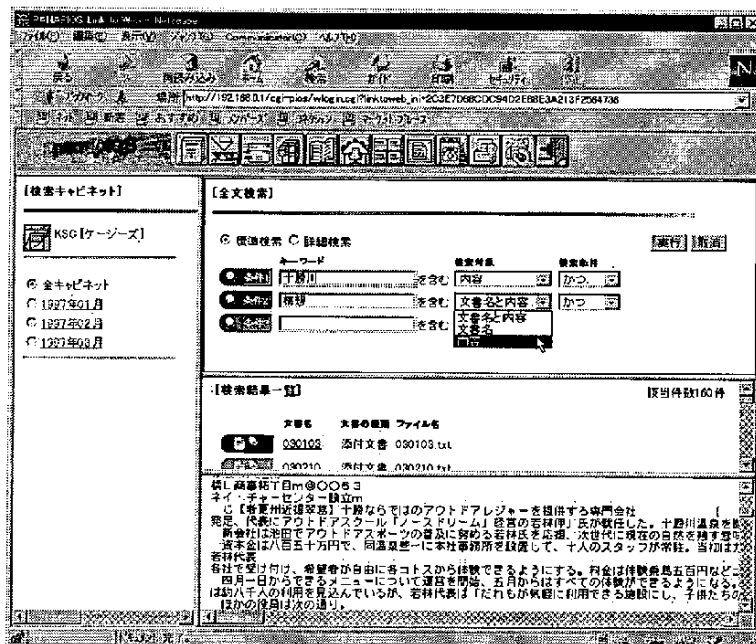
(D) 検索画面が表示されるので、探したい記事の条件1に“十勝川”と入力する



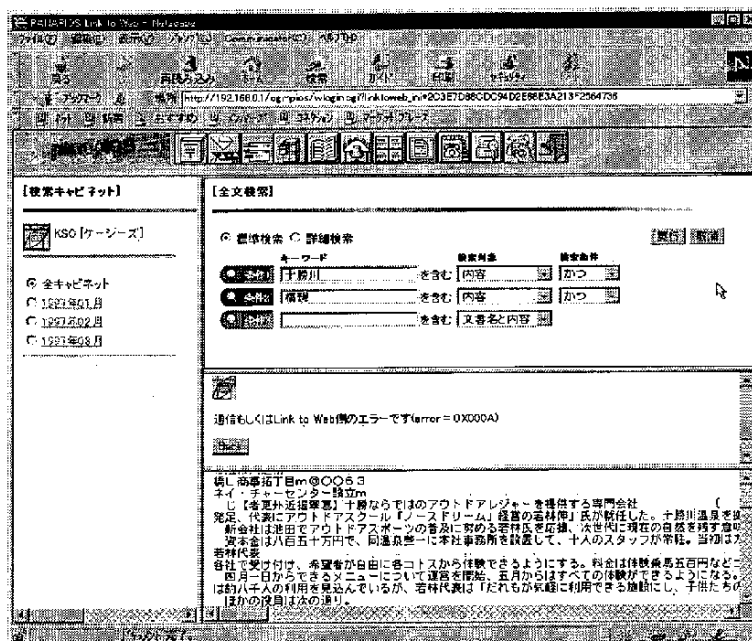
(E) 検索を終了すると、中段にキーワードがヒットした文書名と該当件数、下段には文書内容が表示される。文書中のキーワードはハイライト表示される。



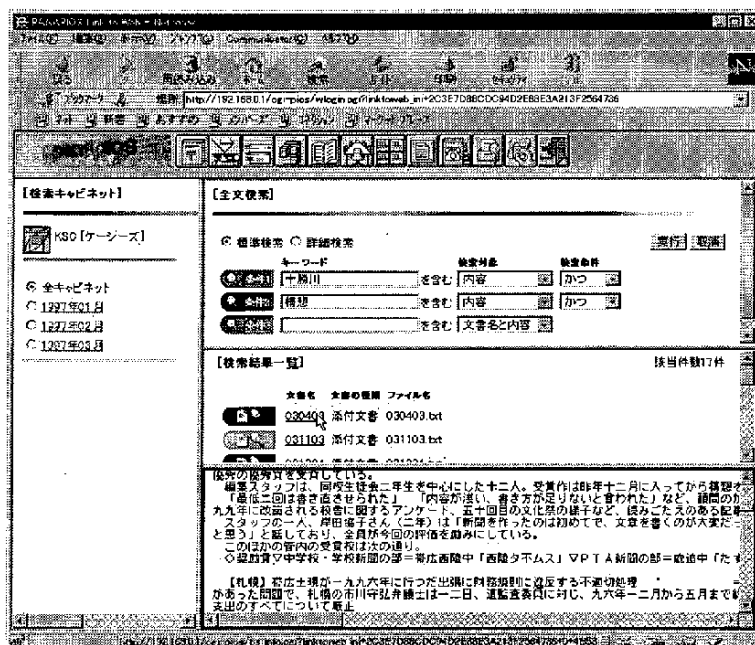
(F) 表示された件数が160件では多すぎるので、条件2に“構想”と入力する。



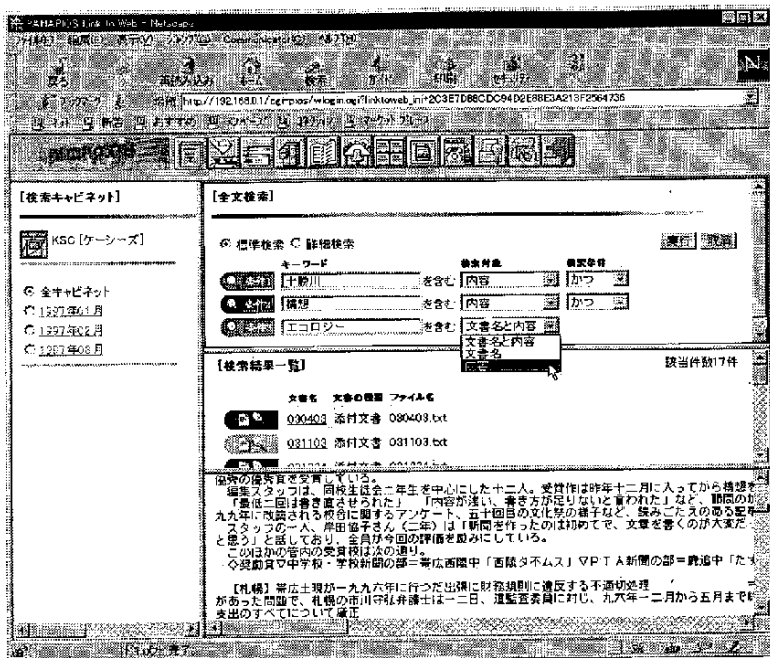
(G) サーバーの処理が間に合わなかったり、ネットワークが混雑していると、このようなエラーメッセージが出る場合もある。



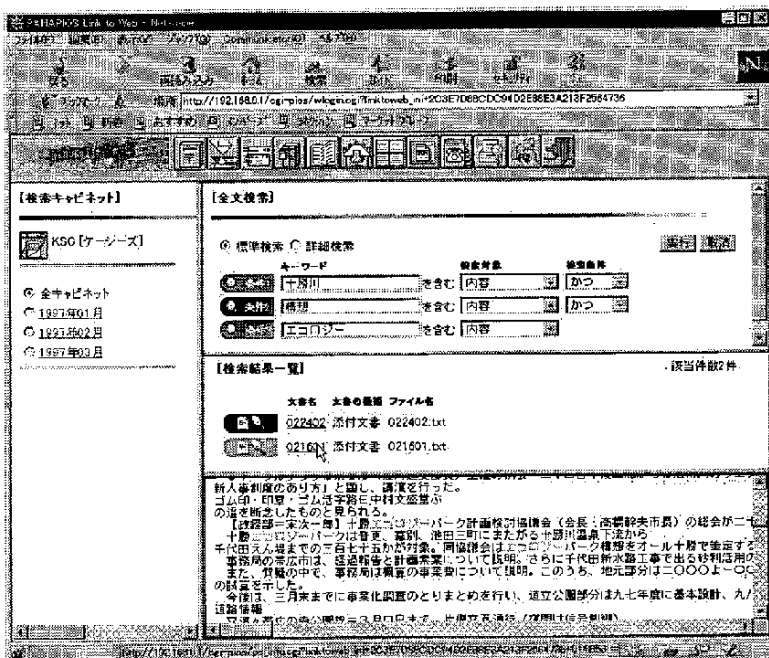
(H) 条件2までの検索結果が出る。該当件数が160件から17件に絞り込まれた。



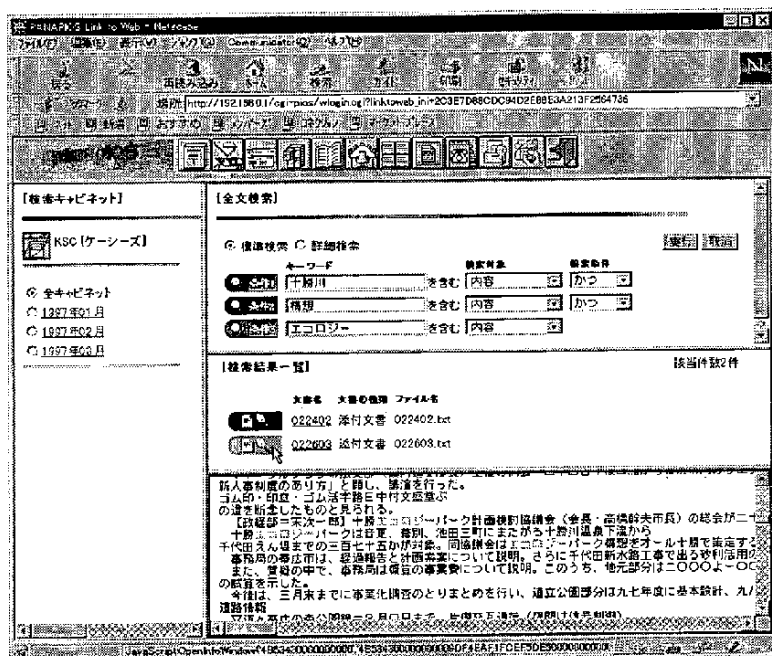
(I) さらに、条件3として“エコロジー”というキーワードを追加する。



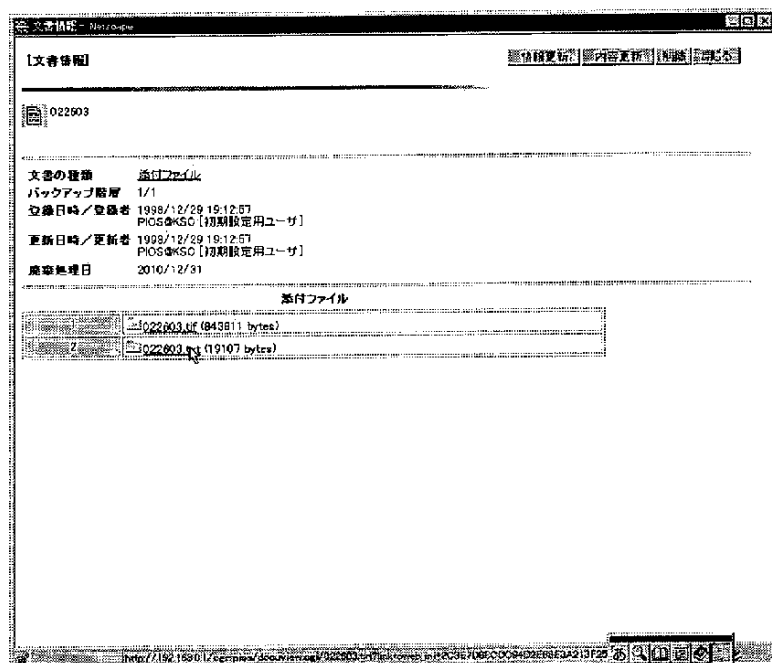
(J) 最終的に2件までに絞り込むことができた。下段の文書表示で閲覧したいページにまちがいないか確認する。



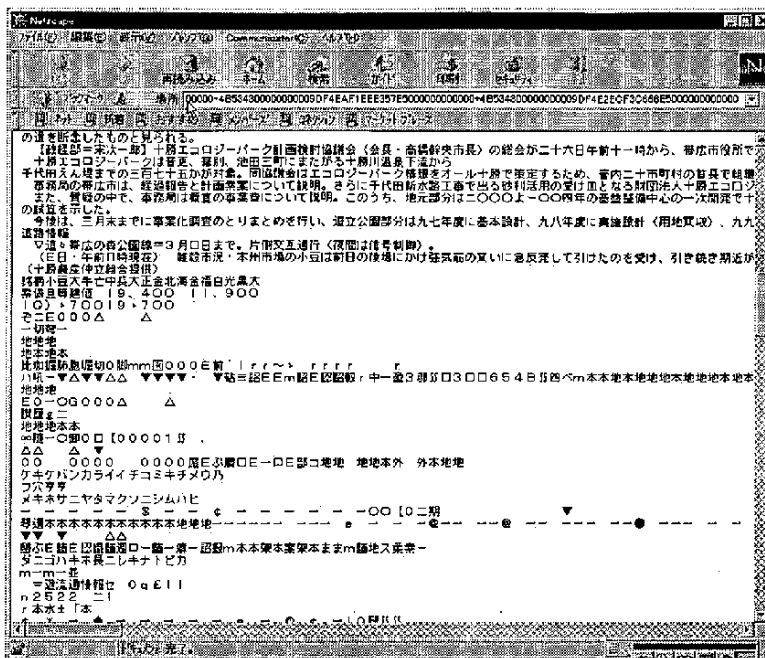
(K) この該当ページを見るためには、中段のアイコン部分をクリックする。



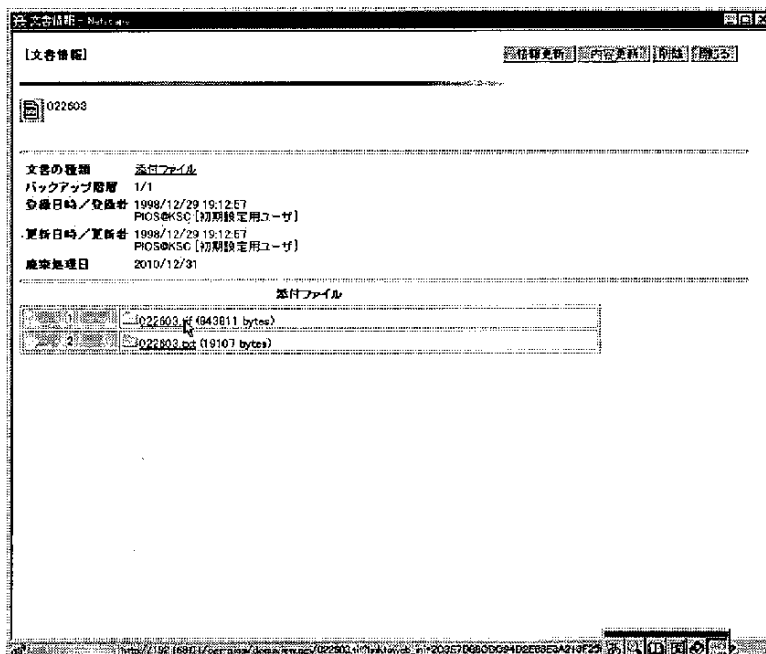
(L) 文書情報の添付ファイルとして、2つのファイルが表示される。1（上）が画像データ、2（下）がテキストデータ。まずテキストを見るために、下のファイル名のところをクリックする。



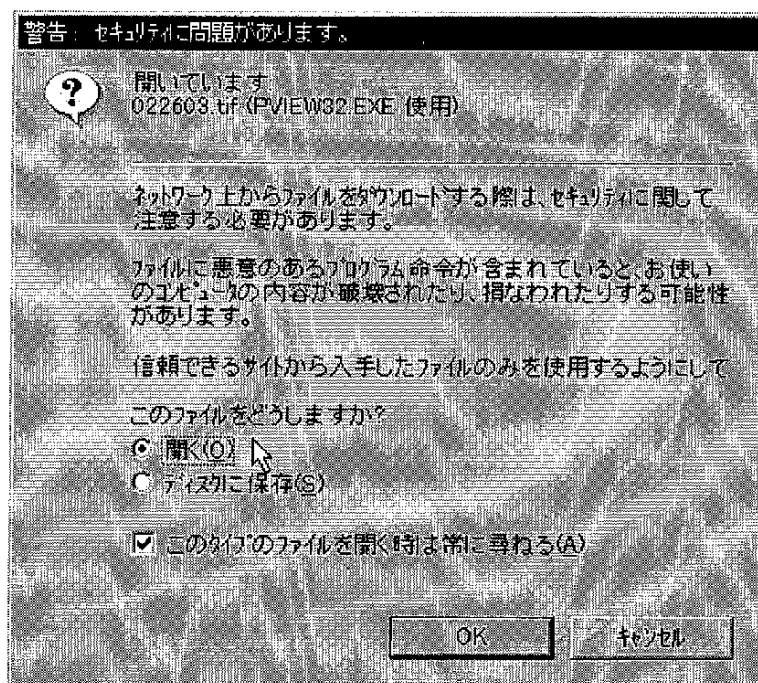
(M) 文書が表示される。テキストデータが必要な時は、ここからフロッピーに保存する。



(N) 次に画像データを見てみる。



- (O) 画像を開く際に、この様な警告が表示される。セキュリティに問題がなければ、“開く”を選んでOKボタンを押す。



- (P) 閲覧用のソフト「パナビューアー 3 2」が起動され、画像データが表示される。



(Q) プラスの虫メガネアイコンを使って拡大する。ここでは、全体表示で拡大率14パーセントになっている。



(R) 画面上で十分読める大きさまで拡大する。この画面では拡大率44パーセント。



(S) ページ全体を印刷する場合は、ファイルメニューの中から、印刷を選ぶ。



2.1.7 データベースと環境

作成したデータベースシステムと、試行利用データベース環境は以下の通り。

(1) 作成したデータベースシステム

(A) データ規模

データベース領域として、画像データサイズの場合約3倍、テキストデータの場合は約10倍の容量を確保しなければならない。

データ種別	仕様	サイズ・単位	総数	必要確保容量
テキストデータ		11KB / Page	2864ファイル	0.3GB
画像データ	400dpi	900KB / Page	2864ファイル	7.6GB
画像データ	200dpi	300KB / Page	5602ファイル	4.9GB
合計				12.8GB

400dpiは1月から3月までを400dpiで登録。それ以降は200dpi。当初は400dpiでスキャンしていたが、データベース容量の都合で200dpiに変換した。しかし、インデックス登録に予想外の時間がかかり、1月から3月までを400dpiで登録した。なお、OCRにかかる都合上、別途に1月から12月のすべてを400dpiでスキャンしている。

(B) 各工程の時間

工程	体制	期間
マイクロフィルムスキャン	1人(30)	8/3～9/18
画像データリネーム	2人(120)	8/5～10/27
リネーム用ツール作成	1人(15)	8/5～10/18
CD-R作成	1人(5)	9/21～10/30
OCR用ツール作成	1人(10)	11/24～12/4
データベース登録用ツール作成	1人(12)	11/26～12/1
データベース登録	1人(5)	12/4～2/9

(2) 試行利用データベース環境 (システム構成など)

システム構成

サーバー (データベースサーバー兼WWWサーバー)

Windows NT 4.0サーバー (Pentium 2 400MHz、128MBS DRAM、HDD 16GB)

クライアント (検索用端末)

Windows 95 (K6-2プロセッサ300MHz、32MBS IMM、HDD 1.2GB) これらの2台を10BASE/TでTCP/IP接続

2.2 問題点

十勝毎日新聞データベースの構築事業を行って、問題の所在が明らかになった。作業レベルで、解決できる問題はその時点で解決したが、いくつかの問題点は解決できなかった。

各作業レベルでの問題点と、全体の問題を、記す。

2.2.1 製作各段階での問題点

(1) デジタル画像化段階

マイクロフィルムのスキャニングと、T I F F化の作業段階で、マンパワーを省略することが、できなかった。

スキャニングしたイメージ画像は、デジタルテキストデータ化の元になるものであり、かつ、全文検索でユーザーが閲覧する画像でもある。データベース構築された画像は、デジタルデータではあるが、見えるカタチはアナログの状態である。この画像の状態が、データベースの性能や仕上がりを大きく左右するだけに、最良の状態に仕上げるのが、求められた。そのために、細心の注意を要するマンパワー、すなわち視認によるチェックが必要だった。T I F F化したイメージ画像が壊れていないか、視認による破損のチェックが欠かせなかった。

画像の中央寄せが必要か、傾いていないか、それらを補正する必要があるかどうか、人間の視認による判断が必要だった。さらに、画像のゴミとりが必要であるか、端っこが黒く汚れている場合、それを白く抜く必要があるか等の、判断も必要だった。

(2) デジタルテキストデータ化段階

(A) O C Rシステムの仕様が非公開のため、新聞紙面の読みとりに、運用しやすいような、カスタマイズができなかった。そのため、回りくどい方法ではあるが、外部から制御する方法をとった。

(B) O C Rのレイアウトの自動認識が完全ではなかった。

文字領域、表領域、写真領域などをダブって認識することがあった。

人為的にレイアウトを指定すれば領域の重複は避けられるが、膨大なデータ処理にマンパワーを投入することは出来なかった。文字の認識率は、現在、市販されているO C Rの中では、100文字中、誤認は1～2文字という、最高度の正確性をそなえていると思われた。

それだけに、レイアウトの自動認識の不完全さが惜しまれた。

(C) しかも、レイアウト領域の順番が、左綴じを前提とし、さらに、横書きを優先しているため、右綴じの縦書き主体の新聞紙面を、読みとるには不都合が多かった。

本格的な新聞データベースを構築する場合は、右綴じ縦読み取りに、O C Rに作り変

えることが、望ましいと思われた。

(D) OCRソフトには、各種の機能が備わっていたが、膨大なデータ容量を扱う新聞紙面用のOCRとして、大容量のデジタル画像を処理できるための、カスタマイズが望まれた。

本格的な新聞データベースを構築する場合、OCRソフトを作成したメーカーと、協議することが望ましいと思われる。

(E) 使用したOCRには、読みとったデジタルテキストデータの一括登録ができなかったため、作業にマンパワーが必要であった。それに伴う人為ミスと時間ロスを解決するために、外部制御のソフトを作成して、連続登録機能自動化を図った。

(F) OCRには、苦手とするものがあった。

個別にあげると、以下の通り。

- ・地紋のある見出しを読み取れない
- ・ヤクモノ（飾り罫線など）が文字として認識されてしまう
- ・横棒を一と誤認識したり、綾杉模様をカッコと誤認識した
- ・原因不明だが、文章認識には得手不得手があった
- ・テレビ欄は得手だったがラジオ欄は不得手だった

こうした苦手や不得手を、OCRのチューンナップで、解決することが、望まれる。

(3) インデックス作成段階

(A) インデックスは、全文検索エンジンの中で、自動処理されている。

そのため、この工程は、自動処理の所要時間も含めて、全文検索エンジンに一任ということになった。

(B) この作業時間が、予測以上にかかり、しかも、いつ終わるかが判断できないという事態になった。こうした状況のため、当事業の後半作業で、スケジュール管理が出来なくなった。

少なくとも、内部自動処理の作業工程の進み具合が表示されていたら、と望まれる。

(C) 表示がなく、ログも残らないので、インデックスの自動処理中に、エラーが発生しても、エラー箇所を特定することができなかった。

(D) バックグラウンドで、インデックスの自動処理を行っているとき、処理済み部分の全文検索を行うとレスポンスが悪くなり、非常に時間がかかった。

(4) 全文検索段階

(A) 当データベース構築中に明らかになった全文検索の問題点は、使用したデータベースエンジンが、小容量のデータベース用に限定しているところに、多くが起因した。

(B) 使用した全文検索エンジンは、階層構造になっていて、会社規模の資料や広報などをデータベース化するには、適しているようであった。

この構造は、新聞記事等の膨大なデータ収納には適していなかった。

(c) ただし、この全文検索は、イントラネットで検索できるようになっているので、分散箇所を結んで利用することも可能であった。このことは新聞記事のデータベースを、インターネットでも利用出来ることを意味し、さらに幅広い活用の可能性を示している。

2.2.2 全体の問題

(1) 新聞という膨大な情報量を、全文検索できるように仕上げるために、なによりも大事だったのは、時間的な問題だった。部分的な改良プログラムの作成や、必要なソフトの開発と同時に、すべての工程が流れ作業となるよう体制を作り上げることが、不可欠だった。この体制を作り上げるまでは、試行錯誤の連続であった。そして、この体制を作り上げることで、大幅な時間短縮が、図れた。全工程を通じて、効率的に事業達成がなされるような、体制を作ることに努力し、成果を上げた。

当事業の、この「大幅時間短縮」体制自体こそ、データベース構築のためのシステムと呼ばれるものである、と考えている。

(2) 使用したソフトには、まだ、バグが残っていた。

そのため、作業が、途中でハングアップするケースがあった。

(3) 膨大な量をあつかうために、サーバーはハイスペックでなければ実用に耐えないことが、わかった。

3 今後の課題

3.1 膨大なデータ量

十勝毎日新聞社に限らず、どこの新聞社も、過去の膨大な記事のデータを持っている。

(1) そのいずれもが、新聞原紙および新聞縮刷版、もしくはマイクロフィルムのカタチで、保存されている。過去の新聞記事を必要とする人は、保存されている上記いずれかの形態を利用して、目的の記事を探し当て、複写するなどの手法をとっている。目的の記事を得るために、利用者は、少なからずの時間と労力を費やしているのが現状である。

(2) この新聞記事が、デジタルデータベース化されれば、新聞記事を利用する際の、時間労力を大幅に省略することができる。さらに、デジタルデータベース化することで、新聞記事を、ネットワークを介した利用法方法へと、発展させることができる。従来の方法とでは、新聞記事の検索はいうにおよばず、探し当てた新聞記事の、プリントアウト、転送などは、比較にならないほど、簡便になる。

(3) 各新聞社とも、そうしたメリットは承知しながらも、過去に蓄えた膨大な新聞記事をデジタルデータベース化することに、二の足を踏んでいる。その理由は、技術的な問題と、それにまつわるコストの問題が、あった。

(4) 当事業の目標は、技術的な問題の所在を明らかにし、コスト計算の試算にも役立つ結果を導き出すことであり、既存の技術、市販のソフトを使用して、新聞記事のデータベース構築を試みたものであった。

3.2 当事業を通じて、問題の所在が明らかになった

これは、今後、新聞記事をデータベース構築する新聞社に共通の問題である。

3.2.1 証明出来たこと

既存の技術、市販のソフトで、問題を含みながらも、新聞記事のデータベース構築は、可能であることが、証明された。しかし、問題は二つあった。

全作業工程の中で、デジタル画像づくりと、インデックス作成に、多量の時間を必要としたことであった。

(1) 一つは、デジタル画像づくりの問題点

(A) 新聞原紙あるいはマイクロフィルム化されたアナログ原稿を、デジタル画像処理をする際に、どうしてもマンパワーが必要であるということであった。

(B) デジタル画像は、T I F F化する対象元でもあり、また、閲覧者が最終的に求めている画像でも、あった。

(C) 精度の高いデジタル画像づくりは、新聞社の校正部のような作業に似ている。校正部は読みやすい紙面づくりのために、多数のマンパワーを必要とする作業現場である。データベース構築作業の中で、デジタル画像づくりは、新聞社の校正部のように、マンパワーが必要だった。

(D) 精度の高い、使いやすいデータベースを追求すると、マンパワーで処理するデジタル画像作りの段階を、重視しなければならなかった。時間短縮のもっとも困難な部分が、この作業段階であった。

(E) とくに古い新聞記事をデータベース化するときは、よりいっそうマンパワーも必要とし、作業時間も今以上に必要となることが予想される。デジタルデータベース化の、大きな課題として残された。

(2) もう一つの問題は、全文検索のインデックス作成に時間がかかる、ということであった。

(A) 全文検索ソフトは、検索時間を短くしようとする、と、検索の仕組みが複雑になるざるをえない。仕組みが複雑になると、そのインデックス作成の時間が長くなる。

全文検索ソフトを制作している各社とも、インデックス作成の時間短縮の努力はしているが、二律背反の性質を負っているがため、作成の時間短縮には成功してはいない。

(B) 新聞記事のような膨大なデータベースともなると、このインデックス作成の所用時間が、よりいっそう大きな問題となった。

(C) ただし、この作業部分は、コンピュータ内部の自動処理であるから、高速コンピュータに置き換えることで、時間短縮を図ることができる。

マンパワーに依存せざるを得ないデジタル画像づくりの作業に比べ、将来の高速コンピュータに期待することが出来る問題ではあった。

(3) この二つ問題を、時間比率で表すと、次のようになる。

データベース構築の全作業で、マイクロフィルム化や、デジタルテキストデータ作りや、個別のソフト作成に要した時間を除いて、問題となる作業が必要とした時間比率は、以下の通りであった。

デジタル画像づくり：インデックス作成 = 3 : 18

作業期間にすると、デジタル画像づくりに、約1人月。

インデックス作成に約6人月。

ただし、このインデックス作成は、当事業の完成締め切りに合わせ、昨年11月から今年2月までの4ヶ月間で打ち切らざるを得なかった。データベース構築は、当初の予定では3ヶ月間の新聞記事を用いる計画であった。デジタル画像も、デジタルテキストデータも、3ヶ月完成済みであった。しかし、インデックスは、3ヶ月分弱（3月28日まで）しか作成できなかったため、構築されたデータベースも、1997年1月から3月28日までの、3ヶ月分弱となった。

3.2.2 その他の問題

作業中の発生した問題点は、市販ソフトを使用しても、外部から制御する方法でプログラムを組めば、ある程度は、解決できる問題であった。もっとも、ソフトを開発したメーカーが、内部公開に応じてくれれば、新聞用に合わせて、もっと簡単に改良ができると思われることも多かった。新聞記事のデータベース化の重要性を鑑み、メーカーの積極的なソフトの内部公開の協力を、求めたいところでもあった。

3.3 新たなる可能性も見えてきた

当事業は、新聞紙面をデジタルデータベース化し、全文検索を可能にすることで、蓄積された新聞記事を、広く活用することを目的としていた。

したがって、完成された成果物は、まず、新聞社での採用・使用が検討された。現時点でも、この成果物は、性能内容に関しては、実用段階にあると判断されている。ちなみに、全面的に記事提供を支援してくれた十勝毎日新聞社では、過去の新聞記事の全文検索システムの実施を、コスト判断による期間設定等、検討されている。

次の、当成果物は、新聞社での使用のほか、公共的な場所での使用が考えられる。現実に、十勝管内の複数図書館に説明提示したところ、導入を積極的に検討したいという回答を得ている。当事業では、各図書館での予算範囲内で、導入できる『十勝毎日新聞データベース』の実施範囲を、こうした公共施設へ、提示する計画である。

3.3.1 可能性の一つはインターネット接続

当事業で使用した、全文検索エンジンが、インターネットでも使えることで、新しい可能性が拓がった。われわれが構築したデータベースは、インターネットを介して、外部から簡単に利用できるように、発展させることが可能である。このことで、閲覧者は、よりいっそうかんたんに、データベースを活用できるようになると、予想される。十勝毎日新聞社では、グループ企業にOCTVという有線テレビ局を擁している。この有線テレビ局が、インターネットのプロバイダー事業に参入することになった。この機会を利し、十勝管内の地域情報流通の活性化のためにも、有線テレビ局および十勝毎日新聞社へ、当事業『十勝毎日新聞記事データベース構築』の全文検索システムの積極的使用を、提案する計画である。

3.3.2 もう一つの可能性

新聞紙面に準じたアナログデータは、すべてデジタルデータ化できる見通しが立ったことである。膨大な量の公文書や、延々と累積された歴史的資料も、当事業の手法を用いれば、デジタルデータベース構築が可能であるという目安がつけられるようになったことである。

3.3.3 終わりにあたって

当事業では、十勝毎日新聞記事を元に、データベース構築を行った。結果としては、まだ、いくつか問題を残してはいるものの、実用に供するデータベースが構築できた。そして、新しい可能性を示すことも、できた。当事業に、改良を加え、継続して、データ量を増やしていくことで、高度情報化社会に不可欠なデータベースになることを、確信して、この報告を終える。

———— 禁無断転載 ————

平成11年3月発行

発行 財団法人 データベース振興センター
東京都港区新橋二丁目13番8号
新橋東和ビル5階
TEL 03 - 3508 - 2430

委託先 株式会社ケーシーズ
帯広市東1条南八丁目2 勝毎ビル4階
TEL 0155 - 25 - 8739

