情 報 学 基 礎 14－1
（１９８９．７．２７）

フルテキスト・データベースの実用化における諸問題
―― 学術情報センターでの事例を踏まえて ――

根 岸 正 光
学術情報センター

論文などの全文をそのまま収録した全文（フルテキスト）データベースは、データベース・サービスの品揃えのひとつとして、多くの情報検索システムの中に加えられるようになってきた。「全文」とはいうものの、図表類は含まれず、本文部分が全文入力されているだけであるが、エンド・ユーザー自身による多様な検索要求に応え得るものとして、期待が集まっている。しかし、現状のシステムを見る限り、機能的に限界があり、利用者側での高度な検索技法を駆使しないと、不要の結果のみが検索されることになりがちである。学術情報センターでは、図表を含む全文データベースの作成を行っており、この経験に即して、全文データベースの作成と検索における問題点を整理し、今後の開発の方向を提示する。

Problems in Construction and Service of Full Text Databases, Discussed with a Case at National Center for Science Information System, Japan

Masamitsu NEGISHI
National Center for Science Information System (NACSIS), Japan
3-29-1 Otsuka, Bunkyo, Tokyo 112, JAPAN

Full text databases have recently become popular as many database services have added the ones to their repertoire. Though full text databases namely contain whole text, and do not usually include figures and tables which appear in the original documents, they are expected to be useful for the various types of searches made directly by end users. However, current systems do not seem to be provided with sufficient functions, and they require users a skill to retrieve relevant records without producing great noises.

NACSIS, an inter-university research institute, has been creating full text databases containing figures and tables. Future developments to be expected for full text databases are proposed by examining the problems experienced there.

1. はじめに

全文データベースは、主として従来の文献抄録データベース、あるいは統計データなどの、数値データベースとの対比において用いられる言葉であって、論文をはじめとする各種文書の全体がデータベース化され、これがオンライン的に検索・出力できるようなものを示すのが通常である。もっとも、昨今はCD-ROMの普及に伴ない、CD-ROM版の全文データベース、すなわちオフライン的利用の全文データベースの方が盛んになるという傾向も見受けられる。一方、これまでの全文データベースは「全文」とはいうものの、本文（テキスト）部分のみをデータベース化したものが殆んどであって、図表を含んだ「完全な」データベースにはなっていない。図表を含めるものはむしろ「マルチメディア・データベース」の呼称のもとに研究開発が進められているが、これは必ずしも論文などの「文書」の、図表を含めた全体をデータベース化することを目指しているわけではなく、従って、全部データベースと称されるべき固有の分野が存在するといってよかろう。それにしても、全文データベースでは何が問題であり、またそれを利用するとどんなよいことがあるのか、といった基本的な点については、必ずしも明らかにされてはいない。

学術情報センターでは、1987年から、化学系学会誌掲載論文の全文データベース化について検討を始め、1989年度から、その利用者向けサービスを開始した。そこで本稿では、同センターでの実際的開発工程での経験を踏まえながら、全文データベースの現状について報告し、今後への展望を得るものとしたい。

2. 全文データベースの位置づけ

本稿で議論の対象とする「全文データベース」について、データベースと称されるもの全体の中での位置づけをまず確認する。「データベース」については、その種別、区分に様々の基準が考えられる。すなわち(1) 文献型／事実（ファクト）型、(2) オンライン提供／オフライン提供、(3) ビジネス向け／研究向け、(4) 文字／数値／画像データ、(5) インハウス利用／公開サービス、などの区分けである。こうした区分はデータベースに対しての各種の視点によっており、論理的にはすべての組合せがありうるのであるが、現実をみると一定の組合せのもののみが多く存在している。

いわゆる情報検索サービスという枠組みの中でいうところの「データベース」とは、その歴史的経過にも依拠して、研究（とくに科学・技術系）向けの文献抄録型（従って文字型データ）データベースのオンライン公開サービスを通常指すことになる。一方「データベース管理システム」（DBMS）との関連では、概ね企業内で維持される（インハウス）事実型・数値データベースを主に対象にしている。[1] 上記両者の中間的なものとして、株価情報、信用情報など、ビジネス用数値データベースの公開サービスがあり、データベース・サービスとしての売上げでは、これが過半を占めるようである。

ところで、全文データベースとは、文献型の延長上にあるが、文献そのままが収録されているという点で、事実型データベースの一種になる。そして、収録の対象文献を学術論文とすれば、それは学術研究向けデータベースになる。一方ビジネス向け全文データベースとして新聞記事データベースがあって、よく利用されるようになっている。また単行書を収容した全文データベース、百科事典の全文データベースなどもある。文献全文を収録するとすれば、テキストのみならず、当然図表、写真などの一切が収容されなければなら

ないが、現状では本文部分だけを収容したものが、「全文」データベースと称してサービスされている。上記のような状況の中で本稿では、学術雑誌論文の、図表をも含む全文データベースのオンライン・サービスを関心の中心にして、その構築、検索などを検討してゆきたい。

3. 新聞記事データベース ―― 抄録型の変型・無停止更新

ビジネス利用向けの全文データベースとしては、まず新聞の全文データベースがあげられる。これは新聞の入稿、編集、版下出力という一連の工程の電算化、CTS（電算写植）化に伴って集積される記事ファイルを、別途データベースとしたものである。データベース化に際して、検索用キーワード、分類などを付与する場合もあるが、いづれにせよその検索は、文献抄録型データベースと同様である。これは、全文とはいうものの一本づつの新聞記事が、分量的にも、また内容的にも、本来抄録的なものであるからである。つまり検索上の方式としては、見出し→本文という一般向きのメニュー方式でもよいし、従来の文献抄録向きのコマンド型検索でもよい。この場合、記事中のすべての語が検索対象となるので、とくに個人名による網羅的検索に、その有用性が発揮されるようである。

新聞記事データベースのサービスでの特徴は、オンライン的無停止更新という運用方式にある。一般の文献型情報検索システムでは、夜間など、サービス運転を停止した後に、バッチ的に新規レコードの追加入力を行なっているが、新聞記事系統では、２４時間無停止であり、また入力も各地の通信社から電話回線を通じて配信されてくる記事を、自動的にデータベースに追加してゆくという方式が特徴になる。検索システムは通信インドーテッド・ファイルを作成するので、そのオンライン的更新には専用のソフトウェアが有効であり、米国ではミニコン上のシステムで全世界にサービスするといった例がある。[2)]

4. 判例データベース ―― 全文データベースのはじめ

全文データベースとして歴史の古いものに、米国の判例データベース・サービスがある。これは判例法主体、弁護士社会という米国の事情によるところが大きく、Mead Data Central社のLEXIS（1973年～）と、West社のWestlawの２大サービスが展開されている。[2,3)] 売上げベースでは、Mead Data Centralは、学術向け文献データベースを主体に350強のデータベースを擁するというDialogに比して、その２倍以上ということであるから、法律まわりの市場の大きさがうかがえる。

ところで、データベースの内容をみると、Westlawの方は、独自の「Shephard索引」と称する判例引用索引を付加している点に特徴があるが、基本的には、両者ともに判決全文そのままをデータベースに流し込み、その全単語から機械的に索引を抽出している。検索システムとしては、論文向けの全文検索システムと同様であって、ことさら特徴的な機能はみられない。すなわち、検索はいわゆる隣接演算によるわけで、用語法などはすべて利用者側でこれを想起して検索する必要がある。これには一種当てもの的要素があり、当面の事象が判決文ではどのように表現されうるか、そのバリエーションを数多く思い付いた人が、よく検索をなしうるという状況にある。

データベースの更新処理からみると、各地の裁判所で出される判決を一刻も早く弁護士に提供する必要があり、この面で新聞記事データベースと似た状況におかれている。そこ

で、全米各地の支所で判決文を取り寄せ、これをOCR入力するといった方式により、判決の数日後にはデータベース化されるよう、新聞記事データベースと同様の無停止更新処理を行っている。

我国では、特定の分野についての判決の要旨を入力したデータベースのサービスが行われているが、判決全文を入力したものについては、その実験が行われたという段階にとどまっている。[4] なお、学術情報センターでは法令のオンライン全文検索システムを開発中であるが、これは箇条書き文書であるので、論文などの全文とは全く異なる。この種のデータベースに対して、利用者側としては、法条間の関連検索の可能性などに期待が向けられるので、むしろネットワーク型のデータベースといった趣がある。

5. テキスト・データベース ―― 完結型データベース

主として文学系統で、古典作品のテキスト全文をコンピュータ・ファイル化したものを、とくに「テキスト・データベース」と称している。我国ではトーマス・マン全集全巻を入力したもの（九大・樋口教授）をはじめとして、欧文、和文の数10点が研究者自身により作成されている。この種のデータベースでは、更新のない完結性が特徴の一つである。利用方法は、語法の統計的研究など、要するにテキスト全文を走査しながら、文字パターンを検出する体のものであるから、一般のオンライン・データベースとはおよそ様子が異なる。昨今はオックスフォード大学で開発されたOCP（Oxford Concordance Program）がパソコンでも動くようになり、欧文系の研究に活用されている。[5] ともあれ、テキスト・データベースは、文学や哲学などにおける全く新たな研究方法を提供していることは事実であろう。

6. ＢＲＳ ―― 学術論文全文データベース

米国BRS Information Technologies社では、1981年から米国化学会との共同で学術論文の全文（本文のみ）データベース・システムを開発し、1983年からこれを一般に提供している。その後米国化学会は独自にサービスを行うようになり、BRSは、現在ではCCML（Comprehensive Core Medical Library）と称する医学雑誌論文全文データベースのサービスで有名である。

BRSでは、雑誌出版社からCTS用磁気テープの供給を受け、これから写植機能コード類を除去すると同時に、著者、標題等のデータ項目種別を判別して、データベース向きに再編集を施すプログラムを、欧米で普及している写植機、数機種に対応して開発した。これにより、CTS→データベースへの変換をほぼ自動的に行うことができるようになっているとのことである。

検索システムは、基本的に隣接演算によるものであって、「ＡとＢが何語以内に連続して現われる文献を検索せよ」といった指令を組み合わせて、検索を行う。こうして検索した結果の表示の際、該当（ヒット）語の出現頻度を表示でき、また該当語の前後のみを抽出して表示するなどの機能により、少ない出力文字数で、結果の当否判断が行われるようにしているが、これが「全文データベース向き」の機能といえるものである。[6]

なお、出版社側での雑誌の売上げ低下に対する危惧を反映してか、全文データベースの方は、雑誌よりも２～３ケ月遅れで提供されているとのことで、当面データベースによる

速報性という観点は阻却されている。また、CCMLの普及には、米国における医療過誤訴訟に対する医師側の防衛措置的側面が強く、ここにも米国の特殊事情が反映されている。こうしたエンド・ユーザー（医師）による直接的検索需要に対応して、1985年からCOLLEAGUEと称するメニュー型検索方式も用意しているが、機能的にはコマンド方式と同一のものである。[7)]

7. ＣＪＡＣＳ ―― 雑誌と全文データベースの一貫生産

米国化学会傘下のCAS（Chemical Abstracts Service）では、同学会発行の論文誌19誌を中心に（CJACS:Chemical Journals of the American Chemical Societyと称する）、さらに英国王立化学会発行誌などを加えた約35誌を対象として、全文データベース（本文のみ）のオンライン・サービスを1985年から提供している(CJO:Chemical Journals Online)。[8)]
CASは雑誌の版元であるから、全文データベースは、雑誌の編集・印刷のCTS化と一体化して運用されている。すなわち、(1) 受け付けられた原稿は電算入力され（数式は別システム）、(2) 棒組み状態で著者校正に廻される（著者は仕上がりページ・イメージを確認できない）、(3) 図表枠を空けた形でCTS出力を得る、(4) 図表は別組みのものを貼込んで、(5) 印刷する、という工程である。ここでは、そもそも(1)の工程で、著者、標題などのタグ付けも行われるので、最初から全文データベースになっているという体裁である。

検索システムは、BRSと概ね同等の機能のものであるが、各単語について文字回転索引（INFORMATION、 NFORMATION/I、 FORMATION/IN、 ･･･ ）を生成して、語の中間一致でも検索できるようにしている。

8. 学術情報センター・化学全文データベース
―― 図表込み全文データベース

学術情報センターでは、1987年から学会誌の全文データベース化について検討を開始し、1989年４月から化学系学会誌の試行サービス（図表出力のためのFAX用電話番号を予め登録しておく）を始め、９月からはこれを本格サービス（FAX番号を出力要求時に指定できる）に移行する。

ここで、データベースの作成方式はBRSと基本的に同様で、学会側から雑誌印刷用のCTS磁気テープの供給を受け、これから写植用機能コード類を削除すると同時に、データベース向きにデータ項目別のタグを挿入するというものである。図表の類は版下の提供を受け、これに図番を付与して光ディスクに焼込む。図表の説明はデータベースの方にも収容されるので、利用者はこれにより図の要否を判別し、見たい図表については、その図番により出力要求を出せば、その図表がFAXで送られてくる。

このデータベース化における問題点の主だったところは次のようである。

(1)写植用ファイルは書体変更の関係もあってか、論文ごとに一本化されておらず、むしろ、標題、著者、引用文献、図キャプションなど、いわば項目種別ごとの横割りになっており、これを論文ごとの縦割りに編成し直す。

(2)数式は特殊文字を含み、高度の入力技能を要するので、別ファイルになっており、その出力を貼り込んで印刷している。

(3)最終校は印刷期限に迫られるため、本体ファイルで修正せず、印画紙上での貼込みで

処置されることが多い。従って、ファイルと雑誌上に現われた文面とでは一部不一致が生じている。

以上の３点は、これまでの印刷工程からみればごく当然のことであるが、データベース化の際にはこれらの処置に相当の「手間」を要する。つまり、プログラムによる自動識別、変換には無理があり、現状では多くの部分について、通常のエディターを用いて人手による編集を行っている。

もっとも(3)については印刷所と協議の上、印刷スケジュールとは別に、最終校正分のファイルへの反映を後追い的に行ってもらうようにした。印刷の都合だけで構成されていた工程をデータベース化の観点も含めて再検討することにより、データベースへの変換処理を効率化しうる点もあるが、何分実務を既定の日程どおり消化しながらのことであるから、改変には相当の時日を要する。

データベースの編集にあたって、ある程度、あるいはかなりデータ内容に立ち入った判断を要する作業として次のものがあげられる。

(1)著者姓名の逆転など正規化。

(2)著者と所属機関の対応。

(3)特殊記号、記法の読み下し（＞Ｃ＝Ｏ → "di-substituted carbony"など）。

(4)キャプションのない図、数式、化学構造式などについてのキャプション補記、図番起こし。

(5)脚注の対応番号の付番。

これらは、そもそも学術論文というものについての一定の知識水準が要求される事項であり、さらに英文誌については英語が読めることも要件となる。従って、実際はかなりのレベルの要員が必要であるが、これを継続的に確保するのはむずかしい。従って、作業マニュアルを整備し、要員を訓練するなど余程の準備が必要であり、そのためにはかなりの期間を要することになる。

９．電子原稿　－－　ＳＧＭＬの可能性

前記化学全文データベースでの事例でわかるとおり、将来的には、著者における原稿作成の段階からデータベース化の観点を含めた編集の方式が望まれる。学術情報センターではこの方向での検討を進めているが、当面の方式としてSGML（Standard Generalized Markup Language）の適用が有力である。[9]

SGMLは、文書の論理構造、すなわち著者、標題等々のデータ項目の種別を計算機プログラムで容易に判別できるよう、タグを挿入しながら、文書を作成するという方式に関わるISO規格である。すでに欧米では各方面に浸透しつつある模様であるが、我が国においては、こうしたものの必要性への認識が高くなかった。この関係では、文書通信を主体に据えたODAの系統の方がむしろよく知られている。[10] 1989年に入って、ECの支援によりヨーロッパで開発されたSGML対応ソフトウェアの我国への紹介があり、これをきっかけににわかに関心が高って、関連業界の協議体も発足するに到った。もっとも、この種国際規格の常として、その日本語化、そしてJIS化が必要であり、これには今しばらくの期間を要すると思われる。

CD-ROMは、オフライン型のデータベースといえるが、この出版は結局既存の出版物のた

めのCTSファイルの変換・再編集という工程によっている。CD-ROM出版の効率化をめざすとすれば、やはり予め文書の論理構造が明示されたファイル（データベース）を作っておく必要があり、ここにSGML普及の一つの要因があろう。

ところで、学界方面をみると、研究者グループでの小規模出版の需要は以前から高く、これにはワープロが多用されている。さらにその高品質化をめざしてDTPへの関心も高く、一部ではTeX普及している。もっともワープロといい、TeXといい、すべて印刷ページの仕上がりイメージに注目したものであって、先にのべたCTSファイルと同等の位置づけにあるといってよい。すなわち、文書の論理構造が明示されたデータベースとは明らかに一線を画している。この際、SGMLやODAにみられるような論理構造→割付構造→印刷出力という段階的な考え方は、文書の再利用、検索などを含む多目的な利用可能性の確保、すなわちデータベース化にとって必須である。

ところで、こうした方向の推進にあたっては、著者における意識の変革が必要になる。印刷物としての論文を唯一の最終目的物と著者が考える限り、全文データベース、あるいはその前提としての文書の論理構造などはおよそ関心の外にあり、データベース業者が勝手にすればよいという程度の認識に落ち着こう。こうした考え方の背景には、全文データベースというものの効用が未だ実感されるに至っていないという事情がある。先にもみたとおり、全文データベースは、現状では米国でも、弁護士のための判例データベース、医師のための医学雑誌データベース、企業人のための新聞データベースなど、職業的裏付けを持つもののみが普及しており、学術研究という観点からの需要は高くないとみられる。そこで、学術研究に向けて、全文データベースがどのように有用たりうるか、全文検索システムの現状と対比させつつ、次に検討してみたい。

１０．全文データベースの効用

学術情報センターでは、先にみた化学全文データベースに先立ち、既製の全文データベースであるHBR（Harvard Business Review）データベースを導入し、これに対する検索システムを開発して、公開した。[11] また、国内での全文データベースの状況について調査を実施した。[12] これらにより得られた知見も総合して、以下のような問題が指摘できる。

（１）保存スペースと遡及検索

雑誌類を手近かに集積しておくには、かなりのスペースと管理事務を要する。また、専門分野をややはずれた周辺的領域の雑誌までは経費的にも購読しにくい。こうした点では、従来図書館がその任務に当たっており、従って文献抄録型データベースで必要な文献が確定できれば、図書館にその原報複写請求を出せばよい。全文データベースはその入手期間を短縮する限りにおいて有用と考えられるが、この目的のためには多種類の雑誌が全文データベースになっている必要がある。CCMLは確かに100種以上の医学雑誌の、文字どおりライブラリーになっており、使いでのあるデータベースといえるであろう。

（２）全文検索機能

全文データベースでは本文全文が収録されているから、検索上極めて有効に思われるが実際には逆効果の方が目立つ。全文であるから、むしろそこには大抵の言葉は使われていて、普通に思い付くキーワードで検索すれば無数の文献がヒットしてしまう。これは語のその文献中での重要度に関わりなく、機械的に検索するから当然であり、結局標題や要旨

などに検索範囲を限定して、所期の文献だけが検索されるようにすることになる。これでは結局文献抄録型データベースと同じことになってしまう。筆者の体験では、国有名詞あるいは特定性の強い語を検索語として検索する際には、確かにノイズが少なく、かつ網羅的検索ができて有効であるが、語の意味が文脈上ではじめて特定されるような場合には、効果的な検索は難しい。この点では、自然文理解に踏み込む以前に、現用の隣接演算をもう少し拡張した機能が考えられてしかるべきものと思われる。

（３）通覧機能と通信速度

現用のシステムは、通信速度の制約から、ヒット部分の限定表示など、全文を見ないで何とか検索結果の可否が判別できるような工夫を施しているが、やはり隔靴掻痒の感は否めない。ページをめくりながら通覧する、あの感覚は他に替え難いものであろう。この点64kbsといった通信速度が前提されれば、全文検索のシステムにも、今とはおよそ異なった局面が拓かれるものと期待される。

１１．結言

本稿では、全文データベースについて、既存の商用サービスを概観し、また学術情報センターにおける経験を踏まえて、その作成工程や検索機能における問題点を指摘した。SGMLへの関心の高まりやISDNによる高速通信の現実化など、全文データベースにとっての環境条件は急速に整いつつあるようにみえる。データベース・サービスの草分け、DialogのR. Summit氏によれば、今後５～１０年の間に、おもな雑誌と新聞はすべてオンラインの全文データベースになるものと確信しうるとのことである。[13] そこで、今後の大きな問題は、著者、編集者、出版社、印刷所、データベース配信者そして読者あるいは利用者などにおける役割意識、価値意識の変革ということであろう。ここでは印刷、出版という、我国においては実に８世紀の百万搭陀羅尼以来の営為に、変容がもたらされることになるからである。

《参考》

1)根岸正光「情報検索システムの構築における利用者の特性とデータベースの特性 － NACSIS-IRの構成を通して」，大学図書館研究，32号，1988，p.1-11.

2)「米国におけるデータベースの現状と展望」，(財)データベース振興センター，1989.

3)"West's Law-Finder － A Legal Research Manual," West Publishing Co., 1988.

4)梅本吉彦「法律情報検索の現段階」，ジュリスト増刊「ネットワーク社会と法」，1988，p.130-135.

5)長瀬真理，西村弘之「コンピュータによる文章解析入門 － OCPへの招待」，オーム社，1986.

6)"BRS Full Text Syllabus," BRS Information Technologies.

7)"BRS Colleague Reference Manual," BRS Information Technologies, 1987.

8)「Chemical Journals Online － Basic User Manual（和訳）」，(社)化学情報協会，1986.

9)Martin Bryan, "SGML: An Author's Guide to the Standard Generalized Markup Language," Addision Wesley, 1988.

10)根岸正光「学術分野における機械可読文書の作成と通信」，学術情報センター紀要，第２号，1989，p.43-52.

11)「Harvard Business Reviewデータベース（NACSIS-IR利用者マニュアル）」，学術情報センター，1989.

12)「全文データベースのシステムに関する現状調査報告書」，学術情報センター，1987.

13)Roger K. Summit, "Outlook for Electronic Information Services," (財)データベース振興センター・データベース国際セミナー，1989，p.21-44.