情報システム　34－2
（１９９１．５．２１）

# 学術情報センターにおける全文データベース検索サービス

原　正一郎　　宮澤　彰　　根岸　正光

学術情報センター

全文データベースは利用者自身による多様な検索要求に応えられるものとして期待が持たれている。学術情報センター（ＮＡＣＳＩＳ）ではデータベース・サービスの一貫として図表を含む全文データベースを公開するとともに、その経験から明らかになったユーザ・インターフェイスの貧弱さあるいは検索ノイズの混入といった全文データベースのかかえる問題点を解決するための研究も行っている。本稿は、このような研究成果を踏まえた「文献の論理構造を考慮した新しい全文データベースシステム」開発への取り組みを、ＮＡＣＳＩＳの学術情報システムの概要と併せて述べたものである。

# Full－text Database Retrieval Servive in NACSIS

Shoichiro HARA, Akira MIYAZAWA and Masamitsu NEGISHI

National Center for Science Information System (NACSIS)

3-29-1 Otsuka, Bunkyo-ku, Tokyo 112, JAPAN

Full text databases are expected to be useful for the various types of searches made directly by users. National Center for Science Information System (NACSIS) has been servicing full-text databases containing figures and tables as its service repertorie, which appeared some problems on the full-text databses, i.e. poor user interfaces and a lot of noises. In parallel with these services, NACSIS has been pursuing the research to resolve these problems. Based on the research, this paper presents our trial of developing "new full text database system reflecting the logical structure of documents". The outline of science information system of NACSIS is also mentioned.

## 1．はじめに

あらゆる学問領域の書籍・雑誌を蓄積し、利用者の要請に応じて適切な文献情報サービスを行うことが図書館、特に研究者に対する大学等図書館の目的である。しかし実際には

1）大学等の図書館に所蔵されている図書・雑誌の所蔵情報が全国規模で整備されていなかったため、学術文献の所在検索・相互貸借が満足に行えないなど、文献資源の有効利用が不可能であった。
2）この問題を解決法として全国規模の学術情報データベースの構築が考えられたが、データベースの対象分野が極めて広範囲であるうえ、各分野の研究動向は専門・深化する傾向があるため、量的効果に依存する商業データベースとしては成立しにくかった。
3）このような経緯から海外の学術データベースへの依存度が高くなった結果、日本は学術情報の輸入超過国となり知識の逆摩擦が生じるにいたり、日本独自の学術情報システムの構築は緊急の課題となった。

学術情報センターは学術情報システムの中枢機関として、全国の大学等の図書館に所蔵されている図書及び雑誌の目録所在情報データベースの構築、社会科学・人文科学・自然科学等の広い領域をカバーする文献情報データベースの構築、ネットワークによる情報ユーティリティの提供を目的として、１９８６年４月に創設された大学共同機関であり、具体的な業務内容は以下のようになっている［1］。

1）学術情報システムの計画・連絡・調整
システムを構成する大学図書館・計算機センター等と連携してシステムの目的を達成するための各種活動の計画・調整を行うとともに、システム内外の関係機関との連絡にあたる。
2）研究・開発
情報図書館学・情報管理学・通信工学・情報科学・システム工学をはじめ、学術情報システムにかかわる人文・社会・自然科学等の諸分野の研究者を広く結集して、総合的な研究開発を実施し、先導性の高い学術研究の需要に対応できるような高度のサービスの実現を可能にする。
3）一次情報に関する目録所在情報データベースの構築とサービスの提供
全国の大学等の図書館における図書・雑誌などの一次情報の体系的・網羅的収集に対応して、これらの目録所在情報のデータベースを、図書館等の協力を得て構築し、迅速・的確な情報提供サービスを実施する。
4）情報検索サービスの提供
書誌・文献抄録などの２次情報データベースや、数値・画像などのファクトデータベースを作成・導入・加工し、人文・社会・自然科学の諸分野の研究者に対し、広く情報検索サービスを提供する。
5）データベース形成の促進
一次・二次情報の総合的な提供拠点として、センター自体により各種のデータベースを形成するほか、学協会・研究者グループによるデータベース形成を支援し、その運用を分担する。
6）情報ユーティリティの提供
テレックス・ファクシミリ・電子メールなどのサービスを光ケーブル・大容量パケット交換機などからなる、デジタル・ネットワークを構築することによって、全国規模で提供する。
7）教育・広報活動の展開
大学等図書館の職員を対象とする実務研修や、研究者等の利用者に対する講習会を実施するとともに、学術情報および学術情報システムに関するシンポジウムを開催するなど、広報活動を展開する。

本稿は、このようなセンター業務のうち全文データベースを例とした情報検索サービスに焦点をあてている。以下では、第２章で学術情報センターのインフラ・ストラクチャーであるネットワーク・計算機資源および情報検索サービスの概要について述べる。第３章では情報検索サービスの具体例として化学全文データベースを取り上げ、現行の全文データベースに関する問題点および解決の方針について整理し、第４章では学術情報センターで検討を進めている新しい全文データベースシステムへの取り組みを紹介する。

## 2．学術情センターの資源

本章では学術情報センターのインフラ・ストラクチャーであるネットワークと計算機資源、およびそ

のもとでサービスされている情報検索システムの概要について述べる。

## ２．１　ネットワーク［2］

学術情報センターのネットワーク（以下では学情ネット）はＣＣＩＴＴ勧告のＸ．２５に基づくパケット交換網で、高速デジタル回線、マルチメディア多重化装置、パケット交換機、パケット多重化装置から構成される。パケット交換機およびマルチメディア多重化装置は学術情報センターといくつかの大学・研究所に設置される。高速デジタル回線はパケット交換の中継線として、マルチメディア多重化装置はパケット交換を効率的に行うために利用される。

データ転送を集中的に管理する目的から、全てのパケットが学術情報センター内のパケット交換機を中継する集中型ネットワークを採用している。高速デジタル回線、パケット交換機およびマルチメディア多重化装置は学術情報センターによって管理されているが、各機関が独自に購入した通信装置を利用することも認められている。また、Ｘ．２５インターフェイスを持たない小規模計算機を学情ネットに接続するため、パケット多重化装置を独自に購入して最寄りのノードに接続することも可能である。

学情ネットはＸ．２５インターフェイスを提供するパケット交換網であるが、ＯＳＩ参照モデルの第４層より上位については固まっていないため、上位層については”ｄｅ　ｆａｃｔ”標準となっているプロトコルやＸ．２５インターフェイスに接続可能なネットワーク方式による論理ネットワークが形成されている。現在、学情ネットで利用されている論理ネットワークには以下のようなものがある。

1）ＸＮＶＴ：大学間コンピュータネットワーク（Ｎ－１）に使用文字種の拡張を行ったＸＮＶＴプロトコルである。これは学情ネットの構築の経緯から、ほとんどの利用者が大学研究者と大学図書館職員で同じ機関に属しているうえ、多くの大学ではＮ－１プロトコルを採用しているためである。応用機能としてはジョブ転送、仮想端末によるリモート・ログインなどがある。
2）ＶＴＳＳ：パソコンなどの小規模コンピュータによるＴＳＳ型接続手順である。ＸＮＶＴおよびＶＴＳＳの場合、利用者コンピュータにユーザ・インターフェイス・プログラム（ＵＩＰ）を実装することにより、グラフィック型マンマシン・インターフェイスを利用することができる。
3）ＭＨＳ(Message Handling System)：大学間・図書館間で形成する電子メール網で、国内電子メールサービス（ＮＡＣＳＩＳ－ＭＡＩＬ）と国際専用回線を利用した米国との国際電子メールサービスがある。ＮＡＣＳＩＳ－ＭＡＩＬはＮＴＴのＤＤＸパケット交換網あるいは公衆電話回線を通じて学術情報センターの電子メールシステム内のメールボックスを介して交換を行う。国際電子メールサービスでは、学術情報センターと米国の国立科学財団（ＮＦＳ）間に設置された国際専用回線を介して、ＮＡＣＳＩＳ－ＭＡＩＬの利用者とＣＳＮＥＴおよびＢＩＴＮＥＴの加入者との間で電子メールの交換を可能としている。
4）ＴＣＰ／ＩＰ：大学ＬＡＮ間の接続に利用されており、応用としてはＴＣＰ／ＩＰの電子メールが主体である。
5）メーカー標準のアーキテクチャ：ＤＥＣｎｅｔやＳＮＡが含まれる
6）Ｇ４－ＦＡＸ：図書館ネットワークの一部に導入されＦＡＸ網を形成している

## ２．２　計算機システム

学術情報センターの計算機システムは図書館目録用・情報検索用および電子メール用の２系統から構成されている。全国の大学等図書館および研究者に対するオンライン目録サービスと情報検索サービスを提供するためのデータベースシステムは、Ｍ－６８４Ｈ／Ｍ－６８２Ｈにより構成されている。このシステムは４台のプロセッサを内蔵するＭ６８４Ｈ（図書館目録用）と２台のプロセッサを内蔵するＭ－６８２（情報検索用）とを疎結合したマルチ・プロセッサ・システムである。さらに内蔵型データベースプロセッサ（ＩＤＰ）により高速のデータベース処理ができる。ファイル・システムの主体となるのは総容量約６４５ＧＢの磁気ディスクであり、さらに半導体記憶装置や光ディスクなどにより様々な種類のデータ蓄積が可能である。

大学研究者・職員に対し国際標準に準拠した電子メール・サービスを提供するための通信システムはＡＣＯＳ１０００を中心に構成されている。また、多量のメッセージを蓄積し、回線を経由して送受信を行うために高速電子ディスク装置と大型の磁気ディスク装置を装備している。

## ２．３　目録データベース

一次情報に関するデータベースとして、学術情報センターでは学術雑誌総合目録データベースの作成サービスとオンライン目録サービスを行っている。欧文編学術雑誌総合目録データベースの編集に当たっては６００機関の協力を得、収録雑誌数約１１万誌、所蔵データ数は約８０万件にのぼっている。ま

た和文編は６２３機関の協力を得て約４万誌、約１００万件が収録されている。この事業は全国の大学等図書館にマークシートを配布して記入を依頼し、当センターにおいて機械可読の形態で入力する方法をとったため、図書館側の理解と協力に依存するところが極めて大きかった。

オンライン目録システムは、各図書館に設置された目録作業専用ワークステーションから学術情報センターの書名・著者名典拠および機械可読目録（ＭＡＲＣ）ファイルを参照することにより、全国規模の書誌調整を可能とするものである。これによって相互貸借システム及び情報検索システムと連携して、一次情報へのアクセスを迅速化することができる。平成３年４月現在でこのオンライン目録システムに参加している機関は１５５にのぼっている。

２．４　情報検索サービス

情報検索サービスはＮＡＣＳＩＳ－ＩＲの名称で実施されている。ＮＡＣＳＩＳ－ＩＲでは学術情報センターが作成・導入・加工したデータベースのみならず、学術情報システムの構成諸機関で作成したデータベースのサービスも行っている。さらに従来の文献情報データベースばかりではなく、数値情報・画像情報などを含む多様なデータベースの提供を目指して、学会はもとより研究者グループによるデータベース作成を積極的に支援している。

現在で学術情報センターが提供している情報検索サービスには海外購入データベース（ＳｃｉＳｅａｒｃｈなど９件）、センター作成データベース（科学研究費補助金研究成果データベース、学位論文データベース、学会発表データベース、学術論文データベースなど１３件）、目録データベース（目録所在データベースなど８件）などがある。

ところで、従来の情報検索システムでは書誌的情報のみのデータベースが扱われていて文献の詳細が不明であるため、検索の適否は最終的には原論文を取り寄せてみなければ判断できず、しかし論文の入手に時間がかかるというジレンマがある。このような従来の書誌情報型データベースに対して、書誌情報を含めた原論文が完全に収録されており、著者の考え方や概念を著者固有の表現で検索でき、検索した文献をその場で見ることのできる全文データベースシステム（以下では全文ＤＢ）に対する要求が高まっている。全文ＤＢサービスは判例やニュース記事のような比較的短い文献を対象に始まったが、計算機の高速化と記憶装置の低廉化にともなって、雑誌などの全文を収録したデータベースサービスも行われるようになった。学術情報センターでもHarvard Business Review、化学系学会誌データベース（学術論文データベース第二系）、電子系学会誌データベース（学術論文データベース第一系）および法令データベース（現行法令データベース）が全文データベースとして公開されている。

以下の章では全文データベースの現状と問題点を整理した後、学術情報センターで検討中の新しい全文データベースシステムについて言及する。

## ３．全文データベースの概要

全文ＤＢは、抄録型データベースあるいはファクト型データベースに対する用語で、雑誌・書籍等の文献全体をデータ対象としたデータベースであり、検索と文献全文の入手がオンラインで可能であるという特徴を持つ。学術情報センターでは１９８７年から国内各学会の協力のもとに、学会誌の全文データベース化についての検討を開始し、１９８９年から化学系学会誌掲載論文の全文データベース（以下では化学全文ＤＢ：正式名称は「学術論文データベース第二系」で）の利用者向けサービスを開始した。

本章では化学全文ＤＢを取り上げ、現行の全文データベースにまつわる問題点および解決の方針について整理する。

３．１　化学全文ＤＢの検索

化学全文ＢＤの利用者は検索コマンドを用いて希望する文献を検索する。検索コマンドは学術情報センターが提供している他の情報検索データベースとほぼ同じである。つまり、一次検索は検索語とフィールドの指定が基本であるが、検索語の指定に際しては前方一致・後方一致・範囲指定などの部分指定や、パラグラフ単位の検索が許される。また、二次検索用は、文献単位の検索としてＡＮＤ・ＯＲ・ＮＯＴ・ＤＩＦＦのブール演算、パラグラフ単位の検索としてＰＡＮＤ（パラグラフ単位での論理積）・ＰＯＲ（パラグラフ単位での論理和）・ＰＤＩＦＦ（パラグラフ単位での論理差）のブール演算、さらにいくつかの補助検索コマンドが用意されている。化学全文ＤＢの結果表示も、基本的には他の情報検索データベースと同様であるが、パラグラフ単位あるいは全文出力といった全文ＤＢならではの機能がある。さらに、化学全文ＤＢの特徴として図表のＦＡＸサービスを挙げることができる。つまり、本文データベース中入っている説明文を参照して、希望する画像ＩＤをＦＡＸ出力コマンドのパラメータと

して指定すると、その図表が利用者のもとへファクスされてくる。このような全文ＤＢに対する利用者側の長所としては以下の事項が指摘されている。

１）全文がオンラインで入手できる。これにより、雑誌類の保管スペースの節約や周辺領域の雑誌購読経費の節約が可能となる。
２）本文をブラウジングできるため、検索の有効性の判定が容易である。
３）網羅的な検索ができる。
４）固有名詞や実験手法などの特定あるいは周辺の情報を利用した検索できる。

一方、欠点としては以下の点が指摘されている。

１）図表の出力にＧ－３ＦＡＸを利用しているため、表示の品位が印刷媒体に比べて著しく劣る。
２）検索ノイズが多い。
３）検索手続きが複雑で、初心者には使いにくい。

出力品位の低さは基本的にハードウェアの問題であり、Ｇ－４ファクスなど高性能の製品が開発・普及するにしたがって解決可能な問題であると考えられる。

これに対して検索ノイズは深刻な問題である。全文データベースでは文献全体が収録されているので検索が容易であるように思われがちであるが、全文であるがゆえに日常的に使用している言葉は殆ど含まれている。したがって、普通に思いつくキーワードで検索を行うと、かなりの量の不適切な文献（ノイズ）にヒットしてしまう。これがノイズの問題である。ノイズが増える原因として、全文ＤＢが従来の情報検索データベースと同様に、文献を単なるフィールドの集合とみなしている点をあげることができる。我々が雑誌を通覧して必要な文献を探す場合、「Ａという用語が章タイトル中に現れていれば探している論文の可能性があり、さらに、その章中にＢというキーワードがあれば、その可能性はさらに高くなる」といったように文献の構造を考慮することで、キーワードの適切な使い分けと検索集合の効率的な絞り込みを行っていると考えられる。このような「文献の構造を考慮した検索法」という視点は、効率的な検索システムの実現において重要であると考えられる。

ところで、現状のシステムではノイズを除去するために、検索を幾重にもかけたり、ヒットした周辺をブラウジングする事によって検索の妥当性を推測している。しかし、検索そのものがコマンドによるものであり、しかもブラウジングの範囲も限られているため、初心者には使いにくいものになっている。ページをめくる要領で全体的に眺めたり、コマンドを意識せずにマウスなどによる直感的操作で検索できるＧＵＩ(Graphical User Interface)の開発が望まれるところである。このような直感的操作を可能にするにも、文献構造を意識した分かりやすいレイアウトを作成する必要があろう。

## ３．２　化学全文ＤＢの作成

化学全文ＤＢの作成は、化学系各学会（高分子学会、日本農芸化学会、日本薬学会）から雑誌印刷用ＣＴＳ（Computer Type Set:電算写植）用磁気ファイルの提供を受けて学術情報センターにおいてなされている。現在、化学系学会誌に限らず多くの学会誌がＣＴＳにより作成されており、このような学会には機械可読な文章ファイルが存在している。しかし、これらのファイルは写植用であるため、データベース化に際して学術情報センターでは以下のようなデータの変換作業を行っている［３］。

１）学会より提供されたＣＴＳファイルから制御コードなどを削除して、平文のテキストファイルに変換する。
２）平文ファイルをエディター上に呼び出し、著者名・所属機関名などの論理項目識別子を挿入して本文データファイルを作成する。
３）全文中からキーワードの抽出を行う。化学全文ＤＢでは日本語のわかち書きと「単語」の抽出にＨＡＰＰＩＮＥＳＳを利用しており、原則としてストップワード（不要語）を除く全単語がキーワードとして抽出される。キーワードとそれに対応したデータベース内のレコード情報は検索用のインバーテッドファイルに蓄積される。
４）図表や数式を、学術情報センターで割当てた画像ＩＤをキーにして光ディスクに焼き込む。ここでは文字データとして処理可能な簡単な表や数式と、画像データとして蓄積すべき図表や数式を仕訳して、後者は画像データベース上の画像ＩＤを割付け、同時に本文中の該当箇所にこのＩＤを挿入するなどの作業を伴う。
５）本文データファイルとインバーテッドファイルをオンライン情報検索システムにロードして利用者に公開する。

データベース作成過程において以下のような問題点が指摘されている。

1）写植機ごとに独自の制御コード体系を持っているので、学会誌ごとに変換プログラムを作成しなければならない。
2）時間上の制約から、最終校正は必ずしも写植ファイルの更新によらず、版下に校正部分を直接貼り込むという便法が行われていることが多い。校正によって生じたファイルと雑誌間の相違部分は、校正原稿を読みながら手作業で修正せねばならない。
3）ＣＴＳファイルはフォントの都合上、論文単位ではなく表題・著者など論理項目別に編集されている場合が多い。そのため、論理項目単位のファイルを論文単位のファイルに再編集する必要がある。
4）論理項目識別子挿入の際には、著者姓名の逆転などの正規化、著者と所属機関との対応、本文と脚注との対応など、編集者の能力が要求される。
5）特殊記号や化学記号の読み下しなど、専門知識が要求される。
6）キャプションのない図、数式、化学構造式などにキャプションと画像ＩＤを補う。

1～3は印刷とデータベース作成における工程差によるものであるが、この差を埋めるには相当の作業量を要する。4～6は学術論文についての編集知識が要求される事項である。実際、データベースの作成においてはかなりの水準の要員が必要であるが、これらの要員を継続的に確保することは困難である。つまり、現状の全文ＤＢの作成法は作業的にも費用的にも効率が低いと言わざる得ない。

文書ファイルの受け取る側で、データ変換作業を効率的に行えるようにするには、印刷物としての書式つまり割付構造(layout structure)情報だけでは不十分で、表題・章・段落など文献の論理構造(logical structure)情報が不可欠である［4］。このような文章構造の国際規格として、ＩＳＯによるＯＤＡ(Official Document Architecture)とＳＧＭＬ(Standard Generalized Markup Language)がある。ＯＤＡは企業などにおける事務用文章ファイルの相互交換を目指すものである。ＯＤＡでは論理構造・割付構造などの機能要素については規定しているが、記述形式は製品開発者に任されている。ただし、ファイル交換の際の形式であるＯＤＩＦ(Office Document Interchange Format)を規定して相互性を確保している。一方、ＳＧＭＬは印刷物の電算ファイルの形態での交換・再編集を目的とした規格と考えられ、ＯＤＡよりも精密な項目設定が可能となっている。もっとも、ＳＧＭＬ自体はシンタクス記述のためのメタ言語であるから、実質的な規格はこれを用いてさらに規定する必要があり、米国ではＡＡＰ(American Association of Publishers)の国内規格などが成立している。

## 4．文献の論理構造を考慮した全文データベースシステムの開発

前章では、ＣＴＳから全文ＤＢへの変換経費の削減、利用者による検索効率の向上およびユーザ・フレンドリーなマンマシン・インターフェイスの作成に当たっては、「文献の論理構造を考慮したシステム開発」が一つの解決法であることを示唆してきた。

文献の論理構造を扱える規格としては前述のようにＳＧＭＬとＯＤＡがあるが、ＯＤＡでは学術出版物を作成する際の記述能力が低く、執筆・データベース化から印刷までの一貫したシステムを構築する上では不十分である。このような理由から、ＯＤＡよりも構造記述能力の高いＳＧＭＬを採用することにした。本章ではＳＧＭＬに依拠したデータベースシステムの開発経過について述べる。

### 4．1　文献の論理構造

文章をイメージとして見ると、文章は字下げ・改行・改ページなどの割付構造によって区別された要素から構成されているように見える。一方、ＳＧＭＬでは章・見出し・段落など要素間の相互的構造を対象としている。本稿でいう文献の論理構造とは、ＧＭＬにおけるこのような相互構造と同義である。つまり意味的にまとまった単位であり、かつ表示上も一定の規則によって相互に区別できる単位を文献の論理要素とし、これらの論理要素の相互関係から構成される構造を文献の論理構造とする［5］。直感的には表題・著者名・要旨・章・章見出し・段落などが論理要素に相当し、これらが組み合わされて文献を構成する。文献の論理構造をこのように考えると、これは「文献」をルートとする木構造で表現できる（図1）。

### 4．2　ＳＧＭＬによる文献の論理構造表現

ＳＧＭＬは１９８６年にＩＳＯ８８７９－１９８６として制定された標準一般化マーク付け言語で、文章の表題・著者・要旨などの書誌データに加えて本文中の章タイトル・節タイトル・段落・文などの

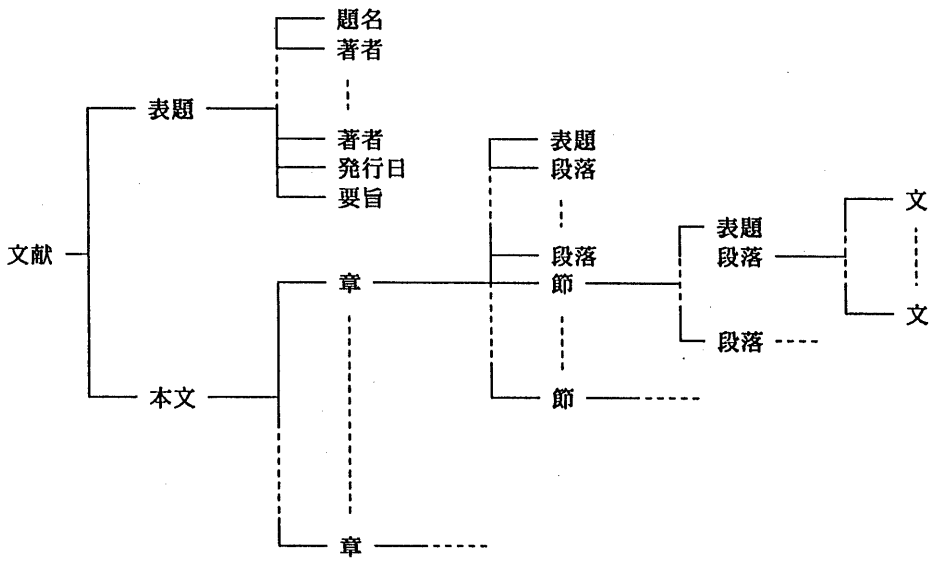


図１　文献の論理構造例

文章の論理構造を表現するための言語規格である。ＳＧＭＬには、１）文書構造のメタ記述、つまり文書構造を記述するための文書型定義機能(DTD:Document Type Definition)、２）ＤＴＤで定義されたタグを用いてマーク付けされた文章を解析する機能、３）省略された文書内容についてのマーク付けの解釈機能、４）文章構造定義の省略機能、５）図表など端末から入力ができない要素を文章中で扱えるようにする要素参照機能、６）文書清書系サポート、７）形式的記述、の機能を持つ。

これらの機能のうち全文ＤＢに関連する機能は

１）文書要素のタグによる識別

２）それによる文書構造の識別

である。つまり、ＳＧＭＬを用いてマーク付けされたテキストデータがあれば、タグを認識する簡単なプログラムを用いることにより、自動的にデータベース用のデータを作成することが可能となる［６］。例えば、図１は次のようなＤＴＤを用いてマーク付けできる。

```
<!ELEMENT  学術文献 ( 表紙(題名, 著者+, 発行日, 要約),
                      本文(章(表題, 段落*, 節(表題, 段落*)*)*)+)  >
<!ELEMENT 題名       (#TEXT)                                      >
<!ELEMENT 著者       (#TEXT)                                      >
<!ELEMENT 発行日     (#TEXT)                                      >
<!ELEMENT 要約       (#TEXT)                                      >
<!ELEMENT 表題       (#TEXT)                                      >
<!ELEMENT 段落       (#TEXT)                                      >
<!ELEMENT 節         (#TEXT)                                      >
```

さらに、このＤＴＤにしたがって本稿のマーク付けを行うと図２のようになる。このように、ＣＴＳ用ファイルがＳＧＭＬ準拠であり、ＤＴＤと文書を同時に受け取れれば、前節で述べたＣＴＳからデータベースを作成する際の問題点のうち、テキストの切り出し・タグの挿入・画像ＩＤの設定についてはほぼ解決できることになる。

一方、ＳＧＭＬのもう一つの機能である文章構造の識別機能を用いることにより検索機能を向上させることが可能である。例えば「<章>の<表題>に"学術情報センター"を含む章で"データ転送"を含んだ<節>の<表題>を取り出せ」という問い合わせが可能となるが、その場合、文書中にＳＧＭＬのＤＴＤによって<章>、<表題>、<節>のタグが定義され、<章>が<表題>と<節>を包含するなどの定義がなされていることが前提となっている。

```
<学術文献><表紙>
            <題名>学術情報センターにおける全文データベース検索サービス</題名>
            <著者>原　正一郎</著者>
            <著者>宮沢　彰</著者>
            <著者>根岸　正光</著者>
            <発行日>1991年5月2日</発行日>
            <要約>全文データベースシステムの.....</要約>
      </表紙>
      <本文> <章>    <表題>はじめに</表題>
                     <段落>図書館の.....</段落>
                     <段落>学術情報センターは.....</段落>
             </章>
             <章>    <表題>学術情センターの資源<表題>
                     <段落>本章では.....</段落>
                           <節>    <表題>ネットワーク</表題>
                                   <段落>学術情報センターのネットワーク...</段落>
                                   <段落>データ転送の集中的管理の観点.....</段落>
                                   .....
                           </節>
                           .....
             </章>
             .....
      </本文>
</学術文献>
```

図2　ＳＧＭＬによるマーク付けの例

しかし、ＳＧＭＬ自身には情報検索機能は用意されていないので、データ操作を行うためには外部の機能を導入する必要があるが、このような問い合わせを十分にサポートできるだけの情報索言語は現状では存在していない。一方、標準として普及しているデータベース用インターフェイスはＳＱＬのみであろう。ＳＱＬはリレーショナルＤＢに対する検索用インターフェイスであり、関係演算による走査対象の指定・挿入・取り出し・更新・削除機能がある。したがって、ＳＱＬは対象を単純な表形式のデータ、つまりＳＧＭＬ形式で表現すれば

<!ELEMENT TABLE((COLUMN_1,....,COLUMN_n)*)>

に限定すれば十分な情報操作能力を持つが、上記の学術文献が持つ「入れ子」構造に対してはまことに不十分である。

### ４．３　検索言語ＤＱＬによる情報検索システムの開発

ＳＱＬは単純な表形式のデータに対しては十分な操作機能を持つが、構造自体の記述能力には限界がある。そこでＳＱＬを拡張してＳＧＭＬにおける文書構造記述子をサポートできるような文書ベース言語ＤＱＬ(Document Query Lanuage)を設計中である。本節ではＤＱＬに概要について述べる。

［文書構造定義］

ＤＱＬではWon Kimらのcomplex object［７］における定義記法をベースとして、ＳＧＭＬのデータ構造記述子をサポートできる記法を採用した。具体的にはＩＳＯ　ＳＱＬ２で規定されている拡張ＢＮＦに準じた記法を用いている。ＤＱＬによる文書構文定義の概要は以下のようになっている。

```
<文書定義>::='CREATE DOCUMENT' <文書型名><文書構造定義>
<文書構造定義>::='('<文書要素構造>')'
<文書要素構造>::=<文書要素名><出現標識> {'TEXT' | <文書構造定義> | <連結子>}
<出現標識>::='?' | '*', | '+' | ' '
            ただし' 'は１回出現する、'?'は０回または１回出現する、'*'は０回以上出現する、'+'
            は１回以上出現することを意味している。
<連結子>::=',' | '|' | '&'
```

ただし、','は順序関係がある、'|'はいづれかが出現する、'&'は順序関係がないことを意味している。

つまりＤＱＬは、ＳＱＬにＳＧＭＬでサポートしている文書構造記述子（グルーピング・出現標識・連結子）を追加したものとみなすこともできる。これにより、

１）文書構造定義において入れ子関係を扱えるようになった。

２）それぞれの要素の出現頻度数を定義できるようになった。

３）要素間の順序に関する定義が可能になった。

という従来の関係データベースにはなかった機能が付与されたため、ＳＧＭＬと同等の文書構造が定義可能となった。この文書構造定義を用いると図１は図３のようになる。

```
CREATE DOCUMENT 学術文献
       (表紙
              (題名 TEXT,
               著者+ TEXT,
               発行日  TEXT,
               要約 TEXT
              ),
               本文
                     (章+
                            (表題 TEXT,
                             段落* TEXT,
                             節*
                                    (表題 TEXT,
                                     段落* TEXT
                                    )
                            )
                     )
       )
```

図３　ＤＱＬによる文書構造定義例

［文書問い合わせ式］

ＤＱＬの問い合わせ式はＳＱＬと類似のキーワードを用いており、その概要は以下の通りである。

```
<文書問い合わせ式>::=<文書問い合わせ指定> <集合演算子><文書問い合わせ指定>
<集合演算子>::='UNION' | 'INTERSECT' | 'EXCEPT'
<文書問い合わせ指定>::=SELSCT<部分文書構造指定>
                       FROM<原文書型指定> AS <部分文書構造指定>
                       WHERE<探索条件>
<部分文書指定>::=--<文書型定義に準ずる>
<原文書型指定>::=<文書型名>
<探索条件>::=--ＳＱＬにおける<探索条件>に準ずる
```

この文書問い合わせ式を用いると、「章の表題に”全文データベース”を含み、その章中の節に”ＳＧＭＬ”と”ＳＱＬ”を含む章」のような例は次のように記述される。

```
SELECT 学術文献
FROM 学術文献
AS (本文.章)
WHERE (章==((SELECT 章
             FROM 学術文献 AS 本文.章.節
             WHERE 節 LIKE '%ＳＧＭＬ%' AND 節 LIKE '%ＳＱＬ%')
             INTERSECT
            (WHERE 表題 LIKE '%全文データベース%')))
```

このように、ＤＱＬは従来の行を基本とした検索操作から構造を持った文書情報を操作しようとしている点に大きな特徴がある。現在ＤＱＬはユーザが出し得る検索要求を想定したシミュレーションを通じて、機能の拡張・整備を進めている。

### ４．４　データベースシステム

全文ＤＢはデータベースの管理と検索を受け持つメインフレームと検索インターフェイスであるワークステーションおよび両者を結ぶ通信系から構成される予定である。

ユーザ・インターフェイスでは、文献構造を図１にような樹系図として端末に上に図形的に表示する。この樹系図の各ノードはオブジェクトであり、検索命令はこれらのオブジェクトに対するメッセージ伝達であるとみなされる。検索は可能なかぎりアイコン化し、操作はマウスで行えるようにして直感的な検索ができるような設計を目指している。さらに、ワークステーションではユーザが作成した図形的検索要求を解析し、メインフレームに対する検索命令をＤＱＬ構文にしたがって組立てて送信するとともに、メインフレームから返送されたＳＧＭＬ形式の文章を解析して表示する機能を持つ。現在、ワークステーションについては、多様な検索例に対するシミュレーションを通じてアイコンの種類・機能および画面レイアウトについての詳細な検討を加えている。

メインフレームでは、ワークステーションから送られてきたＤＱＬ構文による検索命令を解析して文献検索を行う。さらに、検索された全文の中から出力指示に応じた部分を取り出し、ＳＧＭＬのタグとともにワークステーションに返送する。現在、メインフレームについては、ＳＧＭＬパーサの設計とデータベースアーキテクチャの検討を行っている。

メインフレームとワークステーションを接続するネットワークとしてはＬＡＮを用いるが、将来的には遠隔地からのアクセス等を考慮してＩＳＤＮの利用も検討している。

## ５．おわりに

本稿で述べた情報検索システムは、現在の全文ＤＢの持つ問題点に対する一つの提案と試みである。しかし、我々が論文誌の目次をながめ、また論文を通覧してゆくという日常的な方法が小量の文献検索に対して有効であることは日頃経験しているとおりである［８］。そこで大量の論文を収容する全文ＤＢにおいても、このような自然で簡便な方法を通じて必要とする論文を得られるようにすることが研究の目的であり、こうしたシステムの原型が本研究において開発されことを期待している。

本研究は、「文献の論理構造に基づく全文データベース検索システムの研究開発」（科学研究費補助金試験研究（Ｂ））の補助を受けている。またＤＱＬの開発には芝野耕司教授（東京国際大学教授）のご尽力をいただいている。記して感謝いたします。

## 参考文献

［１］猪瀬　博：学術情報システムについて、学術情報センター論文集、和文編第１号、１９８８．
［２］浅野　正一郎、飯田　記子：学術情報ネットワークの現状と展望、ｂｉｔ、Ｖｏｌ．２２，Ｎｏ．２，１９９０，ｐｐ．１２７－１３３．
［３］根岸　正光：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－－、情報処理学会研究報告、Ｖｏｌ．８９、Ｎｏ．６６（８９－ＦＩ－１４－１）、１９８９，ｐｐ．１－９．
［４］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、Ｎｏ．２、１９８９，ｐｐ．４３－５２．
［５］影浦　峡、大山　敬三、宮澤　彰、根岸　正光、鳥居　俊一、絹川　博之：文献の論理構造を考慮した全文検索システム、学術情報センター紀要、Ｎｏ．３、１９９０、ｐｐ．４９－５８．
［６］芝野耕司：ＳＧＭＬと全文データベース、情報処理学会研究報告、Ｖｏｌ．８９、Ｎｏ．６６（８９－ＦＩ－１４－２）、１９８９，ｐｐ．１－８．
［７］WON KIM et.al:"Operation and Implementaion of Complex Objects",IEEE trans. Software Eng., Vol.14, No.7,1988,pp.985-996.
［８］根岸　正光：ＳＧＭＬに依拠する全文データベース・システムの研究開発、学術情報センター・ニュース、Ｎｏ．１５、１９９１、ｐｐ．４－７．