情　報　学　基　礎 33－6
（1994．5．18）

全文検索システムのリーソースとしての ＳＧＭＬ方式データベース

石塚英弘[1)]，伊藤　卓[2)]，千原秀昭[3)]，根岸正光[4)]，中西敦男[5)]，田中洋一[6)]，
[1)]図書館情報大学，[2)]横浜国立大学，[3)]化学情報協会，[4)]学術情報センタ，
[5)]日本化学会，[6)]凸版印刷㈱

SGML方式による全文データベースは、単に各データ項目が明示されているだけでなく、章・節・段落などデータ項目間の階層構造、図表や参照文献と本文の参照関係なども明示している。そのため、これらの特徴を生かして、各種の全文検索ないしハイパーテキスト・システムのデータベースを生成することができる。ここでは、1993年１月号からSGMLに基づいた電子出版になった日本化学会の欧文誌のSGML方式全文データベースを採り上げ、大型コンピュータによるシステムとして、学術情報センターとSTN Internationalの２つの全文検索システム、UNIX WorkstationのX-window上で動くシステムとして、独自開発システムとMediaFinderの２つ、パソコンのシステムとして、Windows 3.1 上のMultimedia Viewerと、Macintosh上のHyperCard等に適用した事例を報告し、SGML方式の全文データベースの有効性を示す。なお、SGMLとはStandard Generalized Markup Languageの略で、1986年にISOの規格ISO 8879となり、1992年にはJIS規格となったものである。

SGML based full-text database as a resource for full-text search system

Hidehiro Ishizuka[1)], Takashi Ito[2)], Hideaki Chihara[3)],
Masamitsu Negishi[4)], Atsuo Nakanishi[5)], Yoichi Tanaka[6)]
[1)]Univ. of Library & Information Science
1-2 Kasuga, Tsukuba-shi, Ibaraki-ken, 305, Japan
[2)]National Univ. of Yokohama,
[3)]Japan Assoc. for International Chemical Information,
[4)]National Center for Science Information Systems,
[5)]Chemical Society of Japan, [6)]Toppan Printing Co.,Ltd.

SGML-based full-text database including a table or a figure (SGML-FT-DB) can be more easily transformed to a database of full-text search or hypertext system than CTS data, since SGML-FT-DB explicitly shows not only a data element but also their hierarchical or referencial relationship. Here, SGML is an acronym of Standard Generalized Markup Language: ISO 8879-1986 or JIS X 4151-1992. We successfully applied SFML-FT-DB of "Bulletin of the Chemical Society of Japan" to 3 types or 6 full-text search or hypertext systems: 2 full-text search systems on main frames: NACSIS-IR and STN International, 2 hypertext systems on workstations: MediaFinder and Toppan's system, and 2 hypertext systems on personal computers: MS Multimedia Viewer and HyperCard.

1．はじめに

電子図書、全文検索、ハイパーテキストなどの発展に伴い、これら検索表示システムのリソースを如何に効率的に作っていくかが問題となり、その解決法としてSGML[1,2]が注目されている。なお、SGMLとはStandard Generalized Markup Languageの略で、1986年に ISOの規格ISO 8879となり、1992年には JIS規格X 4151となったものである。

本稿では、1993年１月号からSGMLに基づいた電子出版になった日本化学会の欧文誌のSGML方式全文データベース[3]を採り上げ、大型コンピュータによるシステムとして、学術情報センターの検索システム(NACSIS-IR)とSTN Internationalの検索システムの２つ、UNIX WorkstationのX-window上で動くシステムとして、独自開発システムとMediaFinderの２つ、パソコンのシステムとして、Windows 3.1 上のMultimedia Viewerと、Macintosh上のHyperCard 等に適用した事例を報告し、SGML方式の全文データベースの有効性を示す。

これまで、検索表示システムの主なリソースは印刷物を作るのに用いた電算写植(CTS) データであった。しかし、CTSデータは印刷のためのデータであって、データベースではない。そのため、データ項目を取り出してデータベース化するための変換プログラムが必要になる。このプログラムはCTSデータに汎用のものではなく、たまたま対象としたデータと検索表示システムを繋ぐ個別のものとなってしまう。またCTSデータは、章・節・段落といった階層構造や、本文から図・表・写真・参照文献・注などへの参照関係は明示していない。階層構造や参照関係をプログラムで完全生成することは困難で、人手による補正が必要となる。このことは、ハイパーテキスト用データベースの作成コストに跳ね返ってくる。なお、CTSデータからハイパーテキストへの変換とその報告としては、たとえば、CACMのハイパーテキスト特集号(Vol.31, No.7)のハイパーテキスト版の作成[4]や、Oxford English Dictionaryのハイパーテキスト版の作成[5]がある。

一方、SGML方式による全文データベースは、単に各データ項目が明示されているだけでなく、章・節・段落などデータ項目間の階層構造、図表や参照文献と本文の参照関係なども明示している。そのため、CTSデータからの変換に見られる問題は生じない。この理由により、SGMLを採用する所が増えている。たとえば、Oxford English Dictionaryも今はSGMLを採用している。

SGMLを検索表示システムに採用した例には、A)検索表示システム提供者、B)DB作成者と検索表示システム提供者の協同プロジェクト、の二つがある。

A の例としては、欧米では、ワークステーション・ベースの電子図書館システムのMercuryやCORE、国際ハイパーテキストシステムWorld Wide Web (WWW)、それにパソコン上のハイパーテキスト・システムGuideなどがある。日本ではNACSIS-IRがある。WWWは独自のデータ入力フォーマット(HTML)を定めているが、これはSGMLで書かれたWWW用のDTD (Document Type Definition)である。また、Guideには、SGML方式で書かれたテキストをGuideの形式に変換するサブシステムIDEX[6]がある。そして、学術情報センターは、大型コンピュータ上の全文データベース検索サービスシステム(NACSIS-IR)のデータ入力として、CTSデータからの変換に加えてSGML方式を採用[7]した。また、SGML方式全文DBを対象とするワークステーション上の検索表示システム[8]も試作した。

B の例としては、Oxford English Dictionaryに関するOxford大学 (DB作成者) とWaterloo大学 (システム開発者) の協同プロジェクト[9]がある。また、アメリカ化学会は前述のCOREにSGML方式の論文誌全文DBを提供することになっている。日本では、学術情報センターのSGML実験誌[10]のCD-ROMと、慶応大学の三田商学研究のCD-ROMがある。

A は検索表示システム側からのアプローチであり、B はあるデータベースについて特定の検索表示システムから見るアプローチである。そこで本研究では、一つのデータベースを、大型コンピュ

ータによるシステム２つ、UNIX WorkstationのX-window上で動くシステム２つ、パソコンのシステム２つ、の３種類計６つのシステムに適用し、問題点を検討した。データベースは、図表を含むこと、学問分野の最先端の事柄についての表記・記述があること、実際に作り続けられており、今後もシステムの検証例に事欠かないことなどを考慮して、日本化学会の欧文誌"Bulletin of the Chemical Society of Japan (BCSJ)"のSGML方式全文データベース（以下、BCSJ全文DBと略す）を採り上げた。なお、SGML方式の学会誌全文DBを作った例としては、情報知識学会誌[1]、学術情報センターのSGML実験誌、慶応大学の三田商学研究などがあるが、いずれも実験的である。

日本化学会は、論文誌全文データベース化のための委員会（委員長は伊藤横浜国大教授で、本報告の著者のうち根岸教授以外は皆その委員）を1990年10月に設置し、慎重な検討準備を経て、欧文誌を1993年１月号からSGMLに基づいた電子出版に切り換えた。これは、SGMLを採用して、図表や写真、それに化学構造式も含む全文DBを作り、このDBからLaTeXによる印刷を行う方式である。

２．ＳＧＭＬ方式全文データベースの特徴

SGMLそれ自身は文書構造を記述する言語である。SGMLを使うことによって、たとえば『本のタイトル』『著者名』『章』『章のタイトル』『節』『節のタイトル』『本文』『パラグラフ』『参照文献』『注』などといった文書の構成要素(element, 以下要素という)を定義するとともに、『章』は『章のタイトル』と『節』で、『節』は『節のタイトル』と『本文』で、『本文』は『パラグラフ』で構成されるといった要素間の階層関係や、『参照文献』や『注』が『本文』中の特定の箇所とリンクしているといった参照関係など、要素相互の構造的関係を定義することができる。この文書構造の定義をSGMLではDTD (Document Type Definition)という。なお、DTDは、図書、論文、マニュアル等、文書のタイプごとに異なるので、タイプごとのDTDはSGMLの個々のアプリケーションとし、SGMLそれ自身は色々な文書構造を記述することが可能な、枠組みの言語とされている。

そして、DTDに従って各要素を示す『マーク』を付けたテキストを、SGML形式のテキストという。『マーク付け』(markup)の形式は、要素の内容を示すテキストを、開始タグ(start-tag) である＜要素名＞と、終了タグ(end-tag) である＜／要素名＞で挟んだものである。要素の中に要素が存在する時は、この形式をネストさせて表現する。

このように文書を構造化することによって、構造を意識した検索が可能となる。たとえば、章のタイトルが"Results"である章に○○○と書いてある論文を検索する、図や表から、あるいは化学構造式や化学反応式から該当する本文の説明を探す、といった検索が可能となる。このような構造の捉え方はハイパーテキストとよく似ている。

なおSGMLでは、通常の文字コードで表現できない記号等の外字は、それに名前を付け、 &名前; としてテキスト内に書き込む。名前はASCII文字で書かれるため、他のシステムとのデータ交換が可能となる。ただし、その外字を印刷あるいは画面表示する場合は、名前に対応する外字のフォントを作って表示する必要がある。

３．ＢＣＳＪ全文データベースの特徴

特徴は次のとおりである。

1)月刊誌で毎号60論文が掲載される大論文誌であること

2)図表を含むこと

・図はドット・データで記述し、別ファイルに納め、本文でそのファイルを参照する

・表はLaTeXで記述し、別ファイルに納め、本文でそのファイルを参照する

3)化学分野特有の表記を含むこと

・H_2O, Ca^{2+}, ^{235}Uなどの上付下付文字が含まれている

・化学構造式や化学反応式の扱い

複雑なものはドット・データとし、図と同じ扱いをする

簡単なものは名前を定義し、DBのテキスト中に &名前; と書き込んである

4)数式はLaTeXで記述

5)DTDはアメリカ化学会で検討中のDTDと互換性があること

我々がBCSJのDTDを設計している時に、米国化学会でも同様の計画があることを化学情報協会を通じて知り、意見交換を行った。その結果、論文著者と所属のリンクの仕方、図表や化学反応チャートと本文とのリンク、参照文献の項目など、日本化学会欧文誌特有の点もあるが、全体としては互換性のあるものとなった。

4．検索表示システムへの適用

前述の特徴を持つBCSJのSGML形式全文DBを、1)大型計算機、2)UNIX workstation (WS)、3)パソコン、3種のハードウェアの上で動く、計6つの検索表示システムに適用した。この3種を選んだ理由は、現在稼働している代表的な種類だからである。表1に、各システムのタイプ、ハードウェアとソフトウェア、各システム用DBの生成方式、適用の状態など、概要をまとめた。以下、各種類ごとに述べる。

4.1.大型計算機上のシステム

NACSIS-IRは、国立の学術情報センターのシステムで、全文DBに力を入れており、大学関係者を中心に多くのユーザを持っているので、採り上げた。また、STN Iternationalは、アメリカのChemical Abstracts Service (CAS)が設立した、文献情報DB、全文DB、化合物DBなどを持つ、世界最大の化学情報サービスである。CASは米国化学会の一部門でもあるが、独立して活動している。

NACSIS-IRは全文検索機能を持った情報検索システムで、図表は大型計算機に接続された光ディスク上に蓄積されており、エンドユーザのコマンドにより、指定したFAXに出力される。現時点ではテキストの上付下付文字は表示できないが、ワードラップ機能はある。

表1　各種検索表示システムへの日本化学会欧文誌SGML方式全文DBの適用

システム名 / 担当者	システムのタイプ[*1]	プラットホーム ハードウェア	プラットホーム ソフトウェア	DB生成方式	適用の状態
NACSIS-IR	全文検索	大型計算機 + FAX[*2]	ORION + UP[*3]	SGML→SGML	運用
STN Intrn.[*4]	全文検索	大型計算機	CAS[*5]独自開発	専用形式に変換	検討中
凸版印刷	hypertext	UNIX workstation	凸版独自開発	SGML→SGML	試作
図情報大[*6]	hypertext	UNIX workstation	SonyのMediaFinder	専用形式に変換	実験
凸版印刷	hypertext	パソコン (DOS)	MS Multimedia Viewer	SGML→RTF[*7]	運用可
図情報大[*6]	hypertext	パソコン (Mac)	HyperCard + UP[*3]	専用形式に変換	試作

[*1] 概略の区分　[*2] 図表はFAXで送る　[*3] user programを付加　[*4] STN Internationalの略
[*5] Chemical Abstracts Service　[*6] 図書館情報大学　[*7] Rich Text Format

検索用DBは、テキストは検索用データベースに、図表はイメージで光ディスクに格納して作成する。従来、全文データベースはCTSデータを変換して作成していたが、1994年に論文誌用の汎用DTDを設定し、SGML方式のDBからは直接変換できるようになった。BCSJの場合は、以前はCTSデータの変換で対応していたが、SGML対応機能が実現した後は、この機能を使ってSGML方式全文DBを日本化学会形式からNACSIS-IRの形式に変換している。NACSIS-IRのDTDは論文誌用の汎用DTDで、BCSJは米国化学会のそれに近いDTDであるが、変換は要素単位に行えるので、問題ない。このことは、SGML方式の長所を示している。

STN Iternationalは、文字データだけでなく、化学構造式も表示できる。ワードラップ機能はあるが、上付下付文字や図表の表示は今後の機能開発に委ねられている。

これまでBCSJは載せたことがないが、欧米の論文誌ではCTSデータ変換方式で検索用DBを生成している。変換プログラムはCASが作ることになるが、SGMLからSGMLへの変換になるかもしれない。技術的には問題がないので、日本側の経費負担など主に経済的な問題が検討がされている。

4.2.UNIX workstation上のシステム

イーサーネット接続のUNIX WSで、X-windowを使ったハイパーテキスト機能を持つシステムに適用した。凸版印刷が開発した論文誌検索表示システム[12]（以下、凸版システムと略記）とSonyのMediaFinderの２つである。両方とも、図表は本文中のボタンをクリックすることにより、別のウィンドウに表示する。また、全文検索機能も持っている。

凸版システムはSGML方式の全文DBを読み込むことができるので、この方法に依った。本文だけでなく、図表へのリンクも含めて自動的にハイパーテキストのDBが生成される。ワードラップもする。

MediaFinderは専用のデータ入力書式を持っており、そこには図表へのリンク情報も書くことができる。そこで、BCSJのSGML方式全文DBをその書式に変換するプログラムを作成し、変換した。

4.3.パソコン上のシステム

Windows 3.1 上で動くMultimedia Viewerと、Macintosh上で動くHyperCardに適用した。

Multimedia Viewerは Microsoftが開発したハイパーテキスト機能と全文検索機能を持つビューアである。ワードラップも行う。画像はテキストと一緒にウィンドウに表示することもできるし、ボタンを使って別のウィンドウに表示することもできる。入力書式としてRTFを用意している。

なお、RTFはRich Text Formatの略で、Microsoftがワープロ文書交換用に開発した規格である。RTFではMicrosoft Wordの書式制御コードがASCII文字で表現されている。また、画像など他のデータへのリンク情報も書くことができる。そのため、他のプログラムがRTFを読んで自分用の書式制御コード付きの文書ファイルに変換できるので、RTFは画像を含む文書の交換用フォーマットとして使うことができる。

本研究では、BCSJのSGML方式全文DBをawkを使ってRTFに変換し、Multimedia Viewer用のDBを作成した。図表、数式、化学構造式等は画像として扱い、テキストと一緒にウィンドウに表示した。また、上付下付文字もフォントを持っていて表示する。

HyperCardはウィンドウでなく、カードを表示単位とし、カード上にテキスト・フィールドやボタンを設定できる。ワードラップ機能もある。画像はカード上に置くことも、ウィンドウとして表示することもできる。また、スクリプト言語HyperTalkでプログラムを書くことによって、独自の

機能を付加することができる。

本研究では、HyperTalkで変換プログラムを書いて、SGMLの要素を該当するテキスト・フィールドに書き込んだ。また、リンク機能もプログラムを書いて実現した。

5．おわりに

以上、検索表示システムのリソースとしてSGML方式全文DBが有用であることを示した。今後は、検索表示システムにSGML方式全文DBの特色を更に生かすこと、使いやすい検索表示機能、すなわち一覧性やノイズの少ない全文検索機能などを検討していくつもりである。

文献

1)ISO 8879-1986, Information Processing - Text and Office System - Standard Generalized Markup Language (SGML), Oct. 15, 1986.

2)C.F.Goldfarb, The SGML Handbook, Oxford University Press, 1990, 664pp.

3a)伊藤卓, 化学と工業, 日本化学会欧文誌の全文データベース化と電子出版化への移行について, 46巻1号, pp.92-95 (1993).

3b)石塚英弘, 伊藤卓, 榎敏明, 千原秀昭, 中西敦男, 田中洋一, 日本化学会欧文誌のSGML形式全文データベースの構築・印刷そして検索, 情報処理学会情報学基礎研究会資料, 29-1 (1993.5).

4)L. Alschuler, Hand-crafted hypertext: Lessons from the ACM experiment, in E. Barrett, The Society of Text, MIT Press, MA, pp.343-361(1989).

5)D.R.Raymond, F.W.Tompa, Hypertext and the Oxford English Dictionary, CACM, 31, pp.871-879(1988).

6)P. Cooke, I. Williams, Design issues in large hypertext systems for technical documentation, in R. McAleese(ed.), Hypertext: Theory into Practice, Ablex, pp.93-104.

7)根岸正光, フルテキスト・データベースの応用動向, 情報処理, 33巻, 4号, pp.413-420(1992).

8)猪瀬博ほか, 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開発研究, 科学研究費研究成果報告書, 学術情報センター, 1993, 158p.

9)G.E.Blake, T.Bray, F.W.Tompa, Shortening the OED: Experience with a grammar-defined database, ACM Trans. Info. Systems, 10(3), pp.213-232(1992).

10)根岸正光, 「SGML実験誌」の出版について, SGML実験誌 1991, pp.i-iii(1991).

11)石塚英弘, 情報知識学会誌, 2巻, 1号, SGML形式による学会誌全文データベースの構築と印刷, pp.23-48(1991).

12)斉藤伸雄, オブジェクト指向による論文誌ビューアの開発, 日本印刷学会年会予稿集 (1993.5).