

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

石塚英弘¹⁾, 伊藤 卓²⁾, 榎 敏明³⁾, 千原秀昭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⁾東京工業大学, ⁴⁾化学情報協会,

⁵⁾日本化学会, ⁶⁾凸版印刷株

日本化学会の欧文誌は1993年1月号からSGMLに基づいた電子出版になった。これは、SGMLに基づいて図表を含む全文データベースを作り、それからLaTeXによる印刷を行う方式である。日本化学会の欧文誌は月刊誌で毎号60論文が掲載される大論文誌であり、化学構造式、反応式、H₂O, Ca²⁺, ²³⁵Uほかの上付き下付き文字など化学物質一般に見られる問題があったが、これを解決した。また、CD-ROMとその検索システムも試作した。本講演では全文データベースの構築、印刷そして、その検索システムについて報告する。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規格となったものである。

Construction, printing and retrieval of
SGML based full-text database of
Bulletin of the Chemical Society of Japan

Hidehiro Ishizuka¹⁾, Takashi Ito²⁾, Toshiaki Enoki³⁾, Hideaki Chihara⁴⁾,
Atsuo Nakanishi⁵⁾, Yoichi Tanaka⁶⁾

¹⁾Univ. of Library & Information Science

1-2 Kasuga, Tsukuba-shi, Ibaraki-ken, 305, Japan

²⁾National Univ. of Yokohama, ³⁾Tokyo Institute of Technology,

⁴⁾Japan Assoc. for International Chemical Information,

⁵⁾Chemical Society of Japan, ⁶⁾Toppan Printing Co., Ltd.

Publication of "Bulletin of the Chemical Society of Japan (BCSJ)" has been changed to the electronic publishing based on SGML since January in 1993. Here, SGML is an acronym of Standard Generalized Markup Language: ISO 8879-1986 or JIS X 4151-1992. This publication is performed as follows; SGML full-text database including a figure and a table is constructed, and then a printed material is generated from the database. CD-ROM and retrieval system for it are also developed. Following problems are solved. BCSJ has a large amount of data; it is a monthly english journal and an issue contains 60 papers. A chemical/reaction formula and a super-/subscript, such as H₂O, Ca²⁺, ²³⁵U, are usually used.

1. はじめに

電子図書ないし電子図書館システムのリソースである全文データベースを如何に構築するか。これは大きな課題である。これまでには、印刷物を作るのに用いた電算写植(CTS)データから作るのが主であった。しかし、電算写植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。すなわち、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。そこで、印刷の後にデータベース化するのではなく、最初からデータベースを構築する手法としてSGML(Standard Generalized Markup Language)が最近注目されている。たとえば、アメリカのカーネギー・メロン大学ではワークステーションをネットワークで結んだ電子図書館システム(Mercury)の開発を進めている。そして、そこにコンピュータ関連学会の雑誌論文を格納しているが、全文データベース作成手法としてSGMLを採用した。また、コーネル大学でも同様のシステム(CORE)を構築しているが、その全文データベースもSGMLを採用することになり、アメリカ化学会が提供することになったという。なお、SGMLは、1986年にISOの規格ISO 8879¹⁾となり、1992年7月にはJIS規格X 4151²⁾となったものである。

一方、我が国では日本化学会が2年半の慎重な検討と準備を経て、欧文誌“Bulletin of the Chemical Society of Japan (BCSJ)”を1993年1月号からSGMLに基づいた電子出版に切り換えた。³⁾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データベースを作り、このデータベースからLaTeX⁴⁾による印刷を行う方式である。また、CD-ROMとその検索システムも試作した。演者は、この欧文誌のプロジェクトのための委員会（委員長伊藤卓横浜国大教授）にSGML方式の経験者として参加した。そこで本講演では、SGML方式による学会誌全文データベース構築について、日本化学会でのSGML形式全文データベースの構築、印刷そして、その検索システムについて報告する。そして最後に、この分野の今後を展望することとしたい。

2. SGML適用事例

まず、SGMLがどの分野で使用されているかを述べ、次いで学会誌への適用について述べる。SGMLが使われている分野は次のとおりである。

1) 社内ドキュメントの作成

IBM社やHewlett-Packard社で実施されている。また、IBM社はSGMLの元となったGML(Generalized Mark-up Language)⁵⁾を開発したことでも知られている。なお、日本では最近、富士通が本格的にやり始めたし、マニュアル制作会社でSGMLに対応する所が出てきた。

2) 政府機関での採用⁶⁻⁹⁾

米国では、CALS(Computer aided Aquisition and Logistics Supportの略、アメリカ国防総省が調達する物資についての電子文書化のプロジェクト)、政府印刷局、標準局、特許庁などがあり、英国では標準局やHMSOがある。また、ECでは議事録のデータベースがSGMLを使って作られている。さらに、ISO自身も規格書案などの作成にSGMLを使うようになっている。

3) 古典テキストの全文データベース

テキスト処理や自然言語処理の研究者の間で、個々に作成した研究材料用の全文データベースを互いに交換・共有しようという動きがあり、そのための組織、Text Encoding Initiativeが1987年に設立された。これまでのところ、全文データベースの対象は著作権の問題がない古典テキストが中心である。

4) 学術書の作成

Bryan の SGML の解説書¹⁰⁾はそれ自身 SGML で作成され、1988年に英国で出版された。また、1989年に米国のマサチューセッツ医学会が SGML 方式による複数メディア出版を始めた。^{6,7)} AIDS 関係の教科書 "The AIDS Knowledge Base" がそれである。SGML を採用した理由は、新しい知見が次々と出てくるのに迅速に対応して改訂するためという。なお、Bryan の SGML の入門書の日本語訳¹¹⁾も SGML で作成され、1991年3月に刊行された。

5) 学会誌・論文誌の作成

2年間の検討と準備を経て、1990年末に情報知識学会（ワーキンググループ主査：石塚英弘）は凸版印刷株の協力により、情報知識学会誌¹²⁾を SGML 方式で出版した。¹³⁾ すなわち、投稿者が作成した電子原稿から学会誌用の SGML 形式全文データベースを構築し、それから学会誌を印刷するシステムを開発した。印刷は SGML にリンクさせた LaTeX を用いて行った。構築した SGML 形式データベースには本文のみでなく、表・図・写真も全て含まれている。

この学会誌には欧文・和文の論文だけでなく、解説や講演記録なども含まれており、また理系のみでなく文系の論文もあるなど多様である。多様性へ対応した点が特徴と言える。また、複雑な SGML のタグを用いずに、通常のワープロを使用して原稿を作ることができる簡易マーク付け法を提案したことも特徴である。

情報知識学会誌より以前の例は今のところ見当たらないが、1991年の時点で検討を行っている学会はあった。たとえば、米国コンピュータ学会(ACM, Association for Computing Machinery)はデータベースシステムと SGML との組み合わせによる新しい出版システムの検討を行い、試験的なシステムを開発している。また、米国化学会は、全文データベースを SGML を使って作成する検討を始めっていた。実際には米国化学会の部局である Chemical Abstracts Service が中心になっていた。

また、以前から学術情報センターは電算写植データから全文データベースを作り、その検索システムを公開していたが、1990年春に凸版印刷株と共に、SGML を使って学会誌を作る研究を開始した。そして、日本で刊行されている理工学系学会の論文誌8誌（高分子学会の欧文誌と和文誌、日本農芸化学会と日本薬学会の欧文誌、電子情報通信学会の欧文誌と和文誌、情報処理学会の欧文誌と和文誌）から計10編の論文を採って、SGML による学会誌（SGML 実験誌）¹⁴⁾を91年3月に作成した。SGML 実験誌で採られた手法は情報知識学会誌のそれと基本的には同じであるが、印刷版だけでなく、CD-ROM 版とその検索システムも作成した点が異なる。なお、この共同研究には図書館情報大学の二教官（石塚英弘・長谷部紀元）も参加した。

一方、慶應義塾大学の三田情報センター（図書館）は大日本印刷株と共同で91年9月に三田商学会の欧文誌と和文誌を SGML 方式で試作した。これは CD-ROM 版で、検索システムも用意されている。

これらは何れも実証実験的な面が強い。情報知識学会誌を除いて単発の試作である。また、情報知識学会誌は第1巻が凸版印刷株、第2巻が大日本印刷株と続いているものの、年一回の発行であり、作業量の点で実証実験的である。

これに対して日本化学会の欧文誌の場合は、月刊誌で毎号60論文が掲載される大論文誌であること、化学構造式、化学反応式、H₂O, Ca²⁺, ²³⁵Uなどの上付き下付き文字など化学特有の問題を持つことなどが挙げられたが、慎重な検討の結果、何れも技術的に解決し、93年1月以降現在まで特に問題なく定常業務として運用されている。その点で、実用化の第一号と言えるだろう。

3. SGML方式の特徴

3.1 SGMLと電子出版

SGMLは単なる印刷ではなく、データベースを中心とする電子出版⁸⁾を念頭において考えられている。従来の出版プロセスが著作・編集・デザイン・印刷であったのに対して、ISO の JTC1 SC18/WG8 の電子出版モデルでは、ワープロ著作、ドキュメント（文書）の構造化、割付指定、ページ記述、印刷となる。このモデルに依れば、構造化の段階で全文データベースができるため、印刷物以外の出版たとえばCD-ROMを、構造化、レイアウト（割付）指定、ページ記述のいずれのレベルからも作りうる。ここで、ドキュメントの構造化がSGML、割付指定がDSSSL(Document Style Semantics and Specification Language, 文書スタイル記述言語, ISO-10179)、ページ記述がSPDL(Standard Page Description Language, 標準ページ記述言語, ISO-10180)に対応する。

ドキュメントの構造はSGMLが表現し、レイアウトはDSSSLが指定する。DSSSLはSGMLで規定されたドキュメントのエレメント(element, 要素)，たとえば『章のタイトル』とか『パラグラフ』に対して、書体、級数（字の大きさ）、字詰め、送りなど文書のスタイルに関する指定を指定する。その後にページ指定となるが、これはSPDLで記述される。SPDLはPostScript類似の言語である。そのため、SGMLの規定に従ったテキスト形式（以下、SGML形式のテキストという）を印刷するにはDSSSLやSPDLの機能を持つものとリンクする必要があるが、SGMLの処理系の中にこれらをリンクする機能が用意されているので、それを使ってリンクする。TeXはDSSSLとSPDLの機能を併せ持つとともに、理工系分野で使われる式や記号の印刷機能も優れている。そのため、SGML形式のテキストの印刷にTeX⁹⁾やそのマクロであるLaTeXが使われることが多い。

一方、CTSにはこのような概念ではなく、写植機に与える命令が記号としてテキストに付加されている。たとえば、章のタイトルには『ゴシック』と『センタリング』『級数』などの命令がたまたま付加されているのであって、『章のタイトル』という認識や、ドキュメントの構造という概念があるわけではない。

3.2 ドキュメント構造

SGMLそれ自身はドキュメント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』などといった個々のオブジェクト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階層的関係や、『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リンク関係など、オブジェクト相互の構造的関係を定義することができる。このドキュメント構造の定義をSGMLではDTD(Document Type Definition)という。また、オブジェクトのことをSGMLではエレメント(element)という。これはデータベースの用語で言えばデータ項目に相当する。なお、DTDは、図書、論文、マニュアルなどといったドキュメント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それ自身は色々なドキュメント構造を記述することが可能な、枠組みとしての言語とされている。

そして、DTDに従って各エレメントを示す『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、エレメントの内容を示すテキストを、開始タグ(start-tag)である<エレメント名>と、終了タグ(end-tag)である</エレメント名>で挟んだものである。エレメントの中にエレメントが存在する時は、この形式をネストさせて表現する。

このようにドキュメント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"である章に〇〇〇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。SGML形式の全文データベースからハイパーテキスト形式の全文データベースを作ることが原理的に可能である。

ドキュメント・タイプごとのDTDはSGMLの個々のアプリケーションなので、SGMLを使うユーザ同士で規約として決めればよい。公開されている規約としては前述のCALSが有名である。また、米国出版協会(AAP, Association of American Publishers)はSGMLによる出版のガイドラインとして、電子原稿の作成とマーク付けに関する標準(Standard for Electronic Manuscript Preparation and Markup)の暫定版^{18, 19, 20, 21, 22)}を作成し、その中で図書(book)、論文(article)、逐次刊行物(serial)のための暫定版DTDを作成して普及を計っている。また、日本でも、日本工業規格『情報交換用電子原稿の記述様式』を作るべく、AAPの規約を元にしてJIS原案が作成された。²³⁾

3.3 テキストの入力

SGML形式のテキストを入力するのは面倒である。なぜなら、通常の原稿には存在しないSGML用タグをエレメントごとに入力しなければならないからである。この負担を軽減する方法として、1) SGML用テキスト・エディタの使用と2)簡易マーク付けの採用がある。

SGML用テキスト・エディタは、DTDに従ってエディタ側でエレメント固有のプロンプトを出す。そのため、ユーザはプロンプトに従って該当するエレメント・データを入れていくだけで、SGML形式のドキュメントが作成できる。この方法は、完成原稿を入力していく時に便利である。

4. 化学会欧文誌へのSGML方式の適用

日本化学会欧文誌の作成にSGML方式を導入する際に考慮された事柄は、次のとおりである。

1)月刊誌で毎号60論文という大量のデータに対応できるか。

- ・学会事務局の対応は可能か
- ・著者に新しい対応を要求するのは困難
- ・印刷会社の対応は

2)化学分野特有の問題を解決しうるか。

- ・化学構造式や化学反応式の扱い
- ・上付き下付き文字
- ・印刷物としての品質が保てるか

3)日本化学会欧文誌用のDTDの設計

4)コスト上昇を招かないこと

検討の結果、上記の4点は次に示すようにして解決された。

1)について検討した結果、SGML導入に伴う新たな作業は学会事務局と印刷会社とで分担できる範囲であると判断し、その分担を決めた。著者に対しては、図表は必ず本文で参照を明示すること、参照文献リストで *idem* や *ibid* は認めず一々書いてもらうことなどを除いて、新しい要求はしないこととした。マーク付けは著者には求めず、学会事務局が対応できる程度のごく限られたものに留めた。電子投稿は今回はせず、検討課題とした。

2)のうち、化学構造式や化学反応式については、複雑なものは図としてドット・データで扱い、

印字	ENTITY	印字	ENTITY	印字	ENTITY	印字	ENTITY
◎	&benz;	◎	&benzf;	-	&sbond;	>	&Ltwsbond;
○	&benzrl;	○	&benzful;	=	&dbond;	<	&Rtwsbond;
○	&benzll;	○	&benzfdl;	≡	&tbond;		

図1 化学構造や化学結合のために用意した記号

印字	ENTITY	印字	ENTITY	印字	ENTITY	印字	ENTITY
←	&tlarr;	→	&trarr;	↔	⇝	↓	⇓
→	&trarr;	→	←	→	→	↑↑	↑
←	⇐	→	←	↔	↭	↑↓	↓
→	⇒	↑	↩	↔	⇆	↔	↕
⇐	⇚	↑	↪	⇒	&lrArr;	↖	↗
⇒	⇛	—	↼	≡	&dblarr;	↙	↘
←	↞	—	⇀	↔	⟷	↖↙	&tswarr;
→	↠	—	↽	↔	⟺	↖↗	↖
←	↢	—	⇁	↑	↑	↑	↶
→	&rarrrtl;	←	⟵	↓	↓	↓	↷
↔	↫	→	&xarr;	↑	↿	○	&tolarr;
↔	↬	↔	⟸	↓	↾	○	&torarr;
↔	⤎	⇒	⟹	↓	⇃	↑	&tlsh;
↔	⤏	↔	⟼	↓	⇂	↑	&trsh;
≡	⇇	↔	⟼	↑	⇑		

図2 化学反応式ほかのために用意した特殊な矢印記号

簡単なものは記号を作つて対応することにした。図1に化学構造や結合の記号を示す。図2には反応式ほかで使用する特殊な矢印記号を示した。ここで、印字とは実際に印字される記号を示し、ENTITYとはその記号をSGML形式全文データベース中で文字列として如何に記述するかを示す。上付き下付き文字の例としては、たとえばH₂O, Ca²⁺, ²³⁵Uなどがある。ここで、H₂O の2はHを、Ca²⁺の2+はCaと共に左隣を修飾しているが、²³⁵Uの235は右隣のUを修飾している。意味的には修飾対象を明示したデータベースとすべきではあるが、手間が掛かりすぎるので、単に上付き下付き文字として扱うこととした。また、印刷物としての品質が保てるか否かについては、問題となりそうなものをテスト・データとして検査することで問題点を把握し、対策を立てることとした。

3)のDTDについては、AAPのDTD、情報知識学会誌のDTD、SGML実験誌のDTDを参考にして検討していたが、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、意見交換を行つた。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データ要素もあるが、全体としては互換性のあるものとなり、将来のデータベース提供も技術的には可能となった。

4)のコストについては、必要な作業を確認した後に見積もりを行い、コスト上昇を招かないことを確認した。

5. 検索システム

検索システムとしては、パソコンのWindows 3.1 上で動くビュアー型のものと、unix workstationのX-Window上で動く検索システムとを試作した。

パソコン上のシステムは印刷物イメージを持っており、表示速度は速い。個人用のシステムと言える。一方、ワークステーション上のシステムは各種のインデックスを持つことができ、ハイパーテキスト的な参照機能も持っている。また、サーバー・クライアント型になっているので、ネットワークを通しての複数ユーザの利用が可能である。このシステムは凸版印刷株で開発されたもので、近いうちに詳しい発表がなされる予定である。

6. おわりに

日本化学会欧文誌のSGML形式全文データベースのプロジェクトは、学術情報センターとも連絡を取っており、センターへのデータベースの提供が決まっている。また、Chemical Abstracts Serviceへの提供も検討されている。一方、学術情報センターは全分野の学会誌を納めうる DTDの検討を進めている。今後、各方面でSGML形式の学会誌全文データベースが構築され、広く世界に公開されていくであろう。

最後に本プロジェクトのためのワーキング・グループのメンバーを敬称略で挙げておく。伊藤卓（横浜国立大学）、千原秀昭（化学情報協会）、榎敏明（東京工業大学）、石塚英弘（図書館情報大学）、中西敦男（日本化学会）、上嶋邦義（同）、大石保夫（同）、田中洋一（凸版印刷株）、高崎進（同）、坂田英俊（同）、佐久間達郎（同）、高柳由美子（同）。

文献

- 1) ISO 8879-1986, Information Processing - Text and Office System - Standard Generalized Markup Language (SGML), Oct. 15, 1986.
- 2) JIS X 4151, 文書記述言語SGML, 日本規格協会, 1992年7月, 231p.
- 3) 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).
- 4) Leslie Lamport, "A Document Preparation System LaTeX—User's Guide & Reference Manual," Addison-Wesley, 1986.
- 5) C. F. Goldfarb, Document Composition Facility Generalized Markup Language: Concepts and Design Guide, IBM Sh20-9188, (1984).
- 6) 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—, 情報処理学会情報学基礎研究会資料, 14-1 (1989.7)
- 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) 山崎俊一, ドキュメント構造記述言語SGMLと電子出版, マルチメディア時代のユーザ・インターフェース (日経コンピュータ別冊 ソフトウエア), 日経BP, 1989年, pp. 229-237.
- 9) 田中洋一, 文書記述言語SGMLとその動向, 情報処理, 32巻, 10号, pp. 1118-1125(1991).
- 10) Martin Bryan, SGML: An Author's Guide to the Standard Generalized Markup Language, Addison Wesley, 1988, 364pp.
- 11) Martin Bryan 著, 山崎俊一監訳, 福島誠訳, 『SGML入門』, アスキー出版局, 1991年3月, 378pp.
- 12) 情報知識学会誌, 1巻, 1号, 情報知識学会, 1990年12月, 98pp.
- 13a) 石塚英弘, SGMLによる情報知識学会誌の編集印刷について, 情報知識学会誌, 1巻, 1号, p. 24(1990)
- 13b) 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).

- 14) SGML 実験誌 1991, 学術情報センター刊, 1991, 79p.
- 15) 根岸正光, 「SGML実験誌」の出版について, SGML実験誌 1991, p.i-iii (1991).
- 16) 根岸正光, 電子原稿・電子出版・電子図書館—「SGML実験誌」の作成実験を通して一, 情報処理学会情報学基礎研究会資料, 24-8 (1991.11).
- 17) Donald E. Knuth, "The TeXbook," Addison-Wesley, 1984. (斎藤信男監修, 驚谷好輝訳, 『TeX ブック』アスキー出版局, 1989).
- 18) Association of American Publishers, Electronic Manuscript Series, Standard for Electronic Manuscript Preparation and Markup", version 2.0, 1987, 162pp.
- 19) Association of American Publishers, Electronic Manuscript Series, Reference Manual on Electronic Manuscript Preparation and Markup", version 2.0, 1987, 136pp.
- 20) Association of American Publishers, Electronic Manuscript Series, Author's Guide to Electronic Manuscript Preparation and Markup", version 2.0, 2nd ed., 1989, 44pp.
- 21) Association of American Publishers, Electronic Manuscript Series, Markup of Tabular Material", version 2.0, 1987, 27pp.
- 22) Association of American Publishers, Electronic Manuscript Series, Markup of Mathematical Formulas", version 2.0, 2nd ed., 1989, 76pp.
- 23) カラーディジタル画像システムの標準化に関する調査研究(データベース) —表記・表現専門委員会平成元年度報告書, 日本電子工業振興協会, 1990, 84pp.