情 報 学 基 礎 14－2
（１９８９．７．２７）

ＳＧＭＬと全文データベース

東京国際大学
芝野耕司

この論文では、SGML の概要を紹介し、従来のデータベースサービスの枠組を述べたのち、全文データベースのインフラとしての SGML の意義を述べる。併せて、データベース管理システムの研究から見た SGML サポート、SGML データベース管理システム実現への課題及データベースサービス全般についての SGML の意義を最近話題のグループウェアとの関連でを述べる。

SGML and Full Text Database

Kohji Shibano

Tokyo International University
1-13-1 Matobakita, Kawagoe, Saitama 350, Japan

In this paper, we first summarize SGML facilities. After discussing current state of database services technologies, we discuss about SGML as an infra structure of full text database services. Further, we review database research activities which relate to support SGML documents. Then we discuss requirements for supporting SGML database. Finally, we discuss impact of SGML based database technology, as an supporting software for groupware, to database srvises.

1. はじめに

SGML は、1986年に国際規格 ISO 8879-1986 [1] として制定された標準一般化マーク付け言語 (Standard Generalized Markup Langauage) の頭字語である。欧米では既に、SGML の利用が電子出版や政府機関での利用を中心に広がっている。日本でも、アメリカ及び EC での特許出願やアメリカ国防総省での採用を契機に遅蒔きながら、最近になって SGML が注目を集めるようになってきた。

通産省では、SGML の普及促進を図るため、日本規格協会に今年度 SGML 懇談会を設け、国内での普及活動を始めた。現在、SGML 懇談会には、約 30 社の企業からなる一般会員と、データベースサービス、印刷業などの関連団体及び学識経験者からなる特別会員が懇談会に参加し、活動を開始した。

この論文では、SGML の概要を紹介し、従来のデータベースサービスの枠組を述べたのち、全文データベースのインフラとしての SGML の意義を述べる。併せて、データベース管理システムの研究から見た SGML サポート、SGML データベース管理システム実現への課題及データベースサービス全般についての SGML の意義を最近話題のグループウェアとの関連でを述べる。

2. SGMLとは？

SGML は、ISO/IEC JTC1/SC18 で規格開発の行われている規格である。SC18 は、テキスト及びオフィスシステムについての規格開発を行っており、SGML は、通常、電子出版用の言語として理解されている。しかし、SGML 自体は、決して電子出版や文書清書系のためのみの言語ではない。

SGML は何を規定し、何に用いられるのかについては、ISO 規格に書かれている定義を参照するのが最も良いであろう。ISO 8879-1986[1] の適用範囲では、SGML 規格の規定範囲を次のように述べている。

(a) 標準一般化マーク付け言語として知られる抽象構文を規定する。この言語は、マーク付けを解釈可能にするための他の情報とともに、文書構造とその他の属性についての記述を表現する。

(b) 抽象構文を特定の文字及び数値に結び付ける参照具体構文、及びいろいろな具体構文を定義するための基準を規定する。

(c) この言語の要素を用いることによって、規格に合致する文書を定義する。

(d) SGML 規格に合致する文書を処理し、文書中のマーク付けの誤りを認識する能力によって規格合致するシステムを定義する。

(e) （画像、グラフィクス又は清書された文書などの）この規格によって定義しないデータを規格に合致する文書に含める方法を定義する。

すなわち、マーク付けに用いるタグの定義とタグが表す文書構造及びタグに付属する属性情報の定義（メタ定義）、マクロ言語で用いる変数と同じような用途及び端末から入力できない文字などを入力するために用いられる参照変数の定義方法、定義されたタグを用いて文書をマーク付けする方法、マーク付けを字句解析する方法、及び外部データを取り入れるための方法を規定している。

メタ情報の定義部は、文書型定義 (Document Type Definition: DTD) と呼ばれ、この機能によって、任意の構造をもつ文書に対して SGML を適用することを可能にしている。SGML での変数は、実体参照 (entity reference) と呼ばれるが、この機能は、例えば、&SQL3 の名前で、“データベース言語 SQL3 作業文書”として、作業中は用い、その後、これを“データベース

言語 SQL3" に置き換えることによって、&SQL3 のすべての参照を置き換えるときなどに用いる。通常の文字集合にない特殊文字を入力するときなどにも用いる。

また、SGML で規定されている字句解析方法では、実際に文書内容部を SGML によってマーク付けされた入力を行う場合に、その入力が簡単になるように SGML タグによるマーク付けの省略時解釈方法を規定している。

一方、SGML の適用範囲としては、次のように規定している。

(a) 異なったテキスト処理言語を用いるシステム間で交換される文書。
(b) 処理が同じテキスト処理言語を用いるとしても、一つ以上の方法で処理される文書。

基本的には、文書を異なったシステム間及び異なったメディア間での文書交換を適用範囲としている。実際、SGML を用いることによって、電子出版システム間での文書交換用に用いることや SGML でマーク付けされた文書をもとに印刷出版、CD-ROM での電子出版及びオンラインネットワークでのサービスを同じ文書データをもとに実現している例が欧米では、既に、幾つも見られる。

SGML 規格は、その記述の難解さで知られている。この SGML 規格の難解さは、一つには、SGML がTeXなどの具体的な文書清書系のための言語ではなく、OSI プロトコルにおける ASN.1 のように抽象構文であることによる。

また、SGML は、文書型定義部で定義されたタグによって文書構造を記述し、その解釈方法を規定しているという意味で、コンパイラコンパイラの一種とも言える。この意味で、SGML 規格書での記述は、通常のプログラム言語の規格書の記述がメタ言語記述であるのに対して、メタメタ言語記述である。

これらの要素を含め SGML には、図 1 に示すさまざまな側面がある。

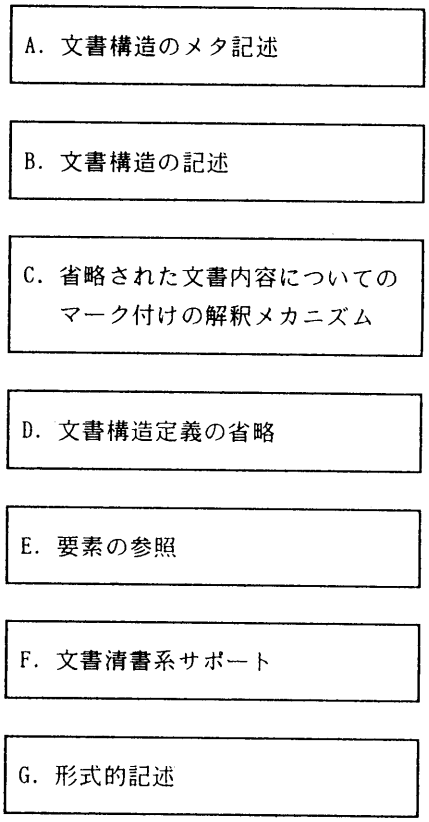

図 1 SGML 規格の構造

SGML の特徴・機能は、図に示す複数の側面の集合として考えられる。

A 及び B は、SGML の最も中心的な機能である文書構造を記述するための文書型定義と文書型定義で定義されたタグを用いてマーク付けされた文書の構文解析を行う機能である。SGML でのコンパイラコンパイラ機能は、静的な構文規則の定義（文書構造の定義）とその構文解析に限られており、通常のコンパイラコンパイラでの実行コード生成に対応する機能は、部分的には、E の参照機能として、実現されていると考えられる。

C 及び D は、データベースにおけるデータ管

理者に当たる文書型定義を行う人がこの定義を簡易に行えるための省略機能及び定義されたタグを用いてのマーク付けを簡単化する又はマーク付けされていない文書を半自動的にマーク付けするための機能である。

E の機能は、文字以外の図形やグラフィクスなどの取り込みや各国文字や記号など端末から入力できない文字を文書中で使用することを可能にする機能である。

F の機能は、実際に SGML 文書を文書清書系の入力とするための補助的な情報を与えるための機能である。

そして、これらの機能が形式的に記述されていることである。

これらの特徴的な機能は、SGML の起源によっていると思われる。

SGML は、ISO 規格のエディタである C. Goldfarb が作った IBM 社の GML を基にしている。IBM 社の GML は、同社の文書清書系である DCF (Document Composition Facility) のマクロ機能を用いたフロントエンドとして実現されたものである[7]。GML においては、陽には文書構造を記述する機能はなく、GML でのマーク付けによる文書構造の記述は、DCF マクロ手続きの実行によって作成される出力イメージから予想されるだけである。

また、GML では、DCF の機能を直接マクロ化して用いていたため、実際の文書作成を行う者は、最小限のマーク付けで期待する出力を得られた。

SGML は、この GML を文書清書系から切り放し、純粋に文書構造を記述する機能に限定する一方、文書構造のメタ定義を行う機能を追加し、省略時解釈機能を強化したものとなっている。

全文データベースを考える際に重要な SGML の文書構造定義機能を表 1 に示す。SGML では、表 1 に示される構造演算子を用いて、実際の文書にタグを用いたマーク付けを行うためのタグの意味を定義する。すなわち、SGML で表現できる文書構造は、表 1 の構造演算子を組み合わせて表現できる構造のクラスとなる。

表 1 SGML の文書構造定義機能

− グルーピング (Grouping)	
(	グループ始め。グループ中の式は他の演算対して一つの単位としてとり扱われる。
)	グループ終り。
− 出現標識 (Occurence Indicator)	
?	選択可能: 0 又は 1 回出現できる。
+	要求されかつ繰り返し可能: 1 回以上出現しなければならない。
*	選択可能かつ繰り返し可能: 0 回以上出現できる。
− 連結子 (Connector)	
,	連結された要素のすべてがモデルグループ中と同じ順番で文書中に現れなければならない。
|	連結された要素の内の一つだけが出現しなければならない。
&	連結された要素のすべてが文書中に出現しなければならない。ただし、どのような順番でもよい。

このように SGML では、自由にタグの名前及び文書型定義としての文書構造の定義を許している。このため、実際に SGML を用いて文

書をマーク付けしようとする場合、あるいは SGML でマーク付けされた文書を交換しようとする場合、もう一歩踏み込んだ規定、すなわち、実際に良く用いられるタグ及び文書構造自体を標準化する必要がある。

これは、文書型定義 (DTD) の標準化としてアメリカ出版協会などで作業が行われている。国内でも、前述の通産省 SGML 懇談会、電子出版協会及び日本電子工業振興協会で、文書型定義の標準化に向けての検討が薦められている。特に、日本電子工業振興協会では、アメリカ出版協会での規格をもとに JIS 化のための検討が行われている[4]。今後、アメリカでの文書型定義の国内への適用可能性、日本語文献のための独自の文書型定義の検討、検討機関間の調整が必要である。

3. 従来のデータベースサービスの枠組

従来の文献データベースの構築では、図 2 に示す作成・利用の流れを前提にしている。

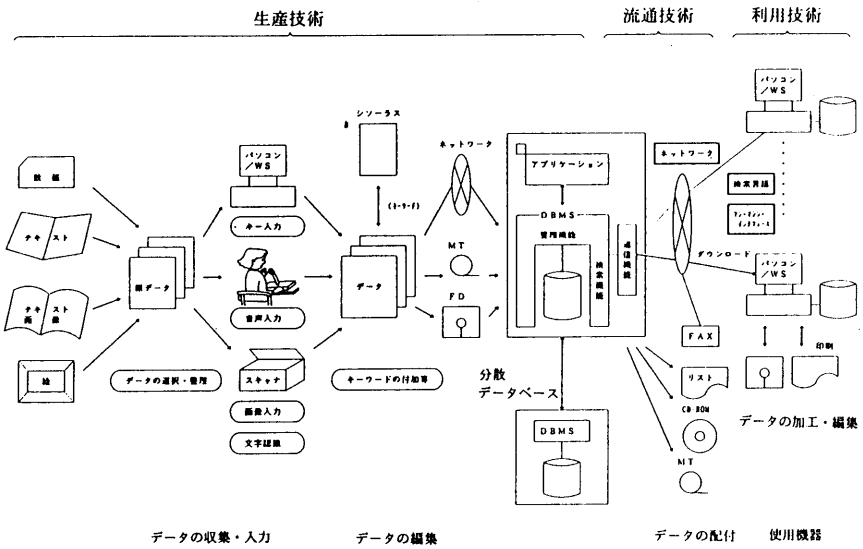

図 2 データベース作成・利用の流れ
([6]の p298 図 10-3-1 からの引用)

図 2 で表される従来の文献検索システムでは、計算機システムから検索される情報は、もともとは計算機システムとは無縁のところで生まれ、存在することを前提にしている。このため、従来の文献データベースの構築では、印刷物として存在する一次情報から書誌情報及び抄録情報を主に計算機システムに投入し、二次情報データベースの構築を行っている。

一方、CTS とワープロの普及によって、最近では身近で目にするほとんどの書類は、ワープロ文書となり、印刷物の多くも、CTS (Computer Type Setting) を用いるようになってきている。

従来の文献データベース構築で仮定されてきた前提は、基本的には、崩れている。しかし、実務上は、現在計算機システムに投入されている情報をもとにデータベースの構築を考えることには大きな障壁が幾つも存在する。

例えば、ISO/DIS 8777[2] で標準化が計られようとしている文献検索システムのための対話型テキスト検索システムコマンドが前提としているデータを既存の CTS あるいはワープロデータから自動的に抽出することは難しい。

対話型テキスト検索システム[2]では、標題、著者名、抄録などのフィールドに対して、文字列の検索を基本とする探索条件の論理式での組

合せを指定して、検索を行う。我々の ISO/IEC JTC1/SC21/WG3 データベース言語ラポータグループでは、1988 年のシドニー会議で、ISO/TC46 から送られてきた ISO/DIS 8777 に対するリエゾンレターに対して、SQL の機能を用いれば、基本的には、簡単に ISO/DIS 8777 で規定している機能を実現でること。SQL2 及び SQL3[3] に日本から提案した各国文字集合サポート機能及び部分文字列の正規表現を用いた照合を可能にする述語を用いれば、ISO/DIS 8777の機能改善を行えることを返答した。

SQL などの対話型の検索コマンドを実現するための基本ソフトへの文献データの投入を考えると、現在計算機システムに存在する CTS やワープロデータなどの機械可読なデータから、ここの列のデータを抽出することの全自動化は難しい。この難しさは、データ中にそのデータについての情報が欠けていることによる。

このことによって、現在のようにほとんどの一次情報が計算機システム中に存在するにも関わらず、これを統合化してデータベース化することには、技術的な困難が存在する。すなわち、一つ一つの一次情報が小箱の中に存在するにも関わらず、この箱を開けて中身をみんなで利用することができないのが現状であり、'この現状の改善には、技術的な課題と社会的な著作権の問題などの課題の解決が必要である。

4. 全文データベースのインフラとしての SGML

前に述べた SGML の機能のうち、全文データベースに関連する機能は、

(1) 文書要素のタグによる識別及び

(2) それによる文書構造の識別

の二つの機能である。

特に、従来の文献検索システムまたはデータベース管理システムとの関連では、SGML を用いてマーク付けされたデータが蓄えられることによって、(1) の機能によって、自動的にシステム中にデータを投入することが可能になる。すなわち、図 2 の生産技術面で SGML は、直接的な影響をデータベース構築に与える。この意味で、SGML は、データベースにとってのインフラとして、基盤技術として極めて重要である。

SGML によってマーク付けされた文書データは、簡単な解釈プログラムによって、自動的にデータベースに投入することができる。SGML の活用は、データベースの生産面での大きな技術革新となる。

しかし、これは、SGML の影響を一つの面だけから見ていることになる。(2) の側面、すなわち、SGML のもつ文書構造の記述能力を生かしてはいない。この側面では、現在のデータベース管理システム及び情報検索システムの機能は明かに不足している。

SGML の文書構造記述を前提にすると、例えば、次のような問合せが可能となる。

> "<図説明>に"SGML"を含み、その<図説明>を含む<図>を含む<章>及びその文書の<標題>を取り出せ"

この例の問合せでは、SGML の文書型定義で<標題>、<章>、<図>及び<図説明>のタグが定義され、<章>が<図>を含み、<図>が<図説明>を含むと定義されていることを前提としている。

従来の形での書誌情報を主とした文献データベースでは、SGML の (1) の機能を中心とした利用で、十分である。しかし、全文データベースの検索では、例えば、ある文字列が本文中にあるかというこれまでのような単純な検索では、欲しい情報を効率的に取り出すことに問題がある。全文データベースの検索においては、この例のように文書構造についての情報を活かした検索によって、より的確に情報を取り出す手段を考慮に入れる必要がある。

SGML の構造記述能力は、全文データベースの検索で必要とされるこのような文書の構造を活用した問合せを可能にする土台となる。

しかしながら、研究を含め、これまでのデータベース管理システム及び情報検索システムでの構造を持った問合せのサポートは、十分とは

言えない。

5. 複合オブジェクトとハイパーテキスト

現在主流である関係データベースのあまりに単純なデータ構造を拡張しようとする試みは、種々なされている。このなかで特に複合オブジェクトをサポートするデータベース管理システムの研究は、SGML がサポートする文書構造のサポートの一部を実現している。

複合オブジェクトのサポートは、System R の拡張の中で最初に提案された[8]。最近の W. Kim 等の研究[9]では、この複合オブジェクトに対する問合せ及びデータ操作を提案している。しかし、これまでの複合オブジェクトに関する研究の文脈中では、関係データベースの列方向での構造のサポートと操作にそのサポートが限定されている。

また、構造記述のサポートの面でも、SGML でサポートしている構造記述演算子のうちグループ化に対応する構造のサポートのみに限定され、出現標識や連結演算子に関連する機能のサポートはない。連結演算子のサポートについては、別の研究で、関係データベースの行の順序をサポートする研究がある。

W. Kim 等の研究では、列に構造を導入したが行は、従来の関係データベースと同じ扱いであり、関係データベースの枠組みをできる限り温存する形での拡張を行っているといえる。グループ化のみのサポートでは、問題があまりでないが、連結演算子等をサポートするとデータ構造の実現値を変更する操作系、例えば、NDL [10]のconnect/disconnectなどの操作が必要になる。これらの機能は、グループ化と順序関係をサポートしているアウトラインプロセッサなどでは、限定された形で実現されている。

ハイパーテキストシステム[5]については、最近注目を集めているが、基本的には、ハイパーテキストシステムは、データベース研究の文脈では、NDL 等のネットワーク型のデータベースと同等の機能を持つシステムであり、その問合せの強力さと柔軟性の面で問題がある。しかし、現在の関係データベースシステムは、あまりの単純化を行い、値のみに依存した問合せを基本にしているため、文書などのマルチメディア情報の取扱については、うまくいかない面が多い。

しかし、単純にネットワーク型に回帰することも、上述のように問題があり、また、実際のハイパーテキストシステムも、その問合せ機能を全面的にユーザインタフェース機能に依存する形で実現しており、データベース機能としてみた場合の問合せ機能は、貧弱であると言える。

6. SGML データベース実現への課題

このように現状で SGML が持つ機能自体をサポートするためにも、幾つかの技術的課題がある。

第一の課題は、SGML の持つ文書構造記述能力を積極的にサポートするデータベース機能の開発である。

第二の課題は、SGML の持つ自由な文書構造のメタ記述能力をサポートすることである。SGML では、文書型定義によって様々な種類の文書の構造を記述することができる。一方、検索の面からみると、この機能は、あまりに多くのメタ情報がデータベース中に投入されることになり、通常の意味での問合せを書くことが困難になり、複数の文書型にわたる問合せの記述量が多くなる。この問題を克服するためには、汎化の概念を問合せ中で利用できる機能が必要になる。

SGML に存在する機能のサポート以外にデータベース機能として考えた場合、補わなければならない機能も幾つか存在する。

SGML では、一つの文書の構造を記述する手段を提供しているが、複数の文書間の関係、特に、整合性制約を記述する手段は、提供していない。データベース機能の開発に当たっては、これらの機能も加える必要がある。

また、SGML にない機能でデータベースにとっ

て最も大切な機能は、データ操作機能である。このデータ操作機能には、NDL 等でみられる構造操作機能をも追加する必要がある。

7. SGMLデータベースと協働的作業環境サポート

SGML データベース機能の実現は、単に、情報検索での全文データベースにとどまらず、より広い応用分野を持つ。この分野としては、グループウェアと呼ばれる協働的作業環境サポートが考えられる。

協働的作業環境サポート機能は、ネットワーク通信サポート、会議サポート、共同執筆サポート等が考えられる[5]。

ネットワーク通信では、Information Lensとして提案されている機能を SGML データベースに対する問合せ機能を用いれば、より柔軟で、強力なものとして実現できるであろう。

共同執筆サポートやグループでのデータの共有の制御は、まさにデータベース管理システムが得意とする分野である。

また、会議サポートにおいても、SGML を基本に拡張されたデータベース管理機能の果たす役割は大きいと考えられる。

8. おわりに

SGML は、データベースの観点から捉えるとすぐに利用できる技術、インフラ等として捉えることもできるが、一方、新たな技術開発の種を数多く提供するものとしても捉えることができる。

SGML から触発される技術課題は、今後の計算機利用の面からも、重要な技術課題が多いと考えられる。

こうした意味で、SGML の積極的な普及を計る一方で、SGML を利用する技術の開発を急がなければならない。

参考文献

[1] ISO 8878-1986, Information Processing - Text and Office System - Standard Generalized Markup Language(SGML), Oct 15, 1986.

[2] ISO/DIS 8777, Documentation - Commands for Interactive Text Searching, Aug 11, 1988.

[3] ISO/IEC JTC1/SC21/WG3 DBL CAN-3, ISO-ANSI (working draft) Database Langauge SQL2 and SQL3, Feb 1989.

[4] 日本電子工業振興協会、カラーデ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） - 情報交換用電子原稿記述様式 -、平成元年 3 月

[5] Irene Greif eds, Computer-Supported Cooperative Work: A Book of Readings. Morgan Kaufman Publishers, 1988.

[6] （財）データベース振興センター編、データベース白書 1989、財団法人 データベース振興センター、平成元年 3 月 31 日

[7] C. F. Goldfarb, Document Composition Facility Generalized Markup Language: Concepts and Design Guide, IBM SH20-9188, 1984.

[8] R. L. Haskin and R. A. Lorie, "On Extending the Function of a Relational Databse Management System," ACM SIGMOD Int. Conf. Management of Data, June 1982.

[9] W. Kim, H. T. Chou, and J. Banerjee, "Operation and Implementation of Complex Objects," IEEE Transaction of Software Engineering, Vol. 14, No. 7, July 1988.

[10] JIS X3004-1987, データベース言語 NDL, 日本規格協会