特別論説

情報処理最前線

情報システム基盤技術としてのSGML†
―文書データベースからWWWそしてCALSまで―

石塚英弘††　根岸正光†††

1. はじめに

欧米に比べて我が国では浸透が遅れていたSGML（Standard Generalized Markup Language）[1,2]も最近急速に様々な分野で注目され，使われるようになってきた．読者の中でも最近SGMLという言葉に接した人は多いのではないか．その場面はCALS（本称は，1985-93年はComputer Aided Acquisition and Logistic Support，その民生化にともない，93年にContinuous Acquisition and Life-cycle Support，94年にはCommerce At Light Speedと変更）[3,4]だろうか，あるいはWWW（World Wide Web）[5]，それとも電子出版だろうか．

SGMLは様々な情報システムでその基盤の技術になると考えられるようになった．たとえば，CALSは新高度情報システムとして期待されているが，その基盤技術としてSGMLが採用されている．また，WWWのデータ入力言語HTML（HyperText Markup Language）はSGMLの一種，より厳密に言えばSGMLの機能の一部を取り入れたもの，である．そして，SGMLは電子出版の基本技術でもある．なぜ，SGMLはこれら一見別に見える分野で使われるのか．その理由は，SGMLがテキストだけでなく図・表・数式等々を含むdocument（以下，文書という）を，データベースのように構造的に記述でき，この技術は分野を問わず適用できるからである．本稿では，SGMLを用いて構造的に記述した電子化文書（electronic document）を「SGML方式文書DB」というが，このSGML方式文書DBが様々な情報システムを改革することになったのである．なお，このDBと一般のデータベースとの相違点など詳細は後で述べることにする．

SGMLについてはすでに解説書[6]～[10]が出版されている．また本誌「情報処理」の解説でも何回か触れられている．たとえば，田中[11]はSGMLをISO規格と電子出版の観点から解説した．根岸[12]はフル・テキスト・データベースの解説の中で，SGMLがフル・テキストDB作成の新技術であることを紹介した．後藤[3]はCALS全般の構成を解説し，SGMLが主要な構成要素の1つであることを示した．もっとも，これらの解説はSGMLを別々の観点から見ているために，読者にとってはかえってSGMLの全貌が掴みにくいかもしれない．

そこで本稿では，SGML方式文書DBが情報システムの基盤となる理由，SGMLとHTMLの違いなどを，SGMLの基本概念に戻って解説することにしたい．そして，SGMLとODAの違いについても簡単に触れる．また最後に最近の話題として，ソフトウェアの分野でのCALSを紹介する．

2. SGMLの趣旨

2.1 文書構造の表現

SGMLの趣旨は文書の構造を表現することである．検索機能や表示機能などはあえて持たず，「餅屋は餅屋」，専門のシステムに委ねている．構造を厳密に表現しておけば，それから各種の検索表示システムあるいは印刷システムの内部フォーマットにプログラムで自動変換できるからである．また，構造が記述されていれば，構造を手掛かりとした検索，たとえば，章または節のタイトルに「WWW」という単語が含まれている章または節の中で，「マルチメディア」という単語が含まれている段落を検索することも可能となる．こ

† SGML as an Fundamental Technology for an Information System — Document Database, WWW and CALS — by Hidehiro ISHIZUKA(Faculty of Library and Information Science,University of Library and Information Science) and Masamitsu NEGISHI (Research and Development Division, National Center for Science Information Systems).

†† 図書館情報大学図書館情報学部
††† 学術情報センター研究開発部

うすれば，たんに「マルチメディア」で全文検索するよりもノイズの少ない検索ができる．この手法は，操作マニュアルや仕様書の内容を更新する際に該当箇所を探す方法としても有効である．

さらに，文書構造を記述しておくことにより，異なるハード・ソフト環境でも使える文書DBが構築できるし，使用していたハードやソフトが時代遅れとなり，別のハード・ソフトに乗り換えることになっても元のデータベースは生き続けられることになる．ハード・ソフト環境によらないから，交換・流通に適しており，また保存形式としても適しているのである．

SGMLでは，文書の構造を表現するために，まず文書構造を定義する．この文書構造定義をSGMLでは，DTD（Document Type Definition，文書型定義）という．DTDでは文書の構成要素の名前と互いの構造上の関係，すなわち，章，節といった階層関係，また図や表の参照といった参照関係，を定義する．そして，DTDに従って実際のタイトル，著者名，テキスト，図，表，式等々を，generalized markup（汎用マーク付け）という記法を使って書く．

なお，SGML自体はDTDを書くための言語である．もしデータベースの用語に例えれば，DTDはスキーマに，SGMLはDDLに対応する．実際の文書のテキストはDTDに従ってマーク付けすることになるから，DTDは文書の記述の仕方を書いたものといえる．その点からいえば，SGMLは文書記述のメタ言語になる．また，SGMLはISO規格で，JIS規格でもあるが，これはDTDを書くための言語の文法を規格にしたのであって，文書構造そのものを規格化したのではない．

ご承知のように，文書によりその構造は異なる．たとえば，単行本は章節で構成されているのに対し，雑誌は各種の記事の集まりとして構成されている．そこで，単行本用のDTD，雑誌用のDTDなど，文書の種類に応じたDTDを用意する．また，構成要素も様々である．論文の場合は抄録が付くが，単行本には付かない．図，表，写真，参照文献等は多くの文書に共通に存在するが，数式，定理の証明，化学構造式，文章の引用等は扱う内容によって存在したり，しなかったりする．もっとも，これまでに様々な機関で色々なDTDが作られてきているので，それを使えば自分で作る必要はない．また，すでに存在するDTDを一部書き換えてもよいし，最初から書いてもよい．なお，前にCALSはSGMLを採用したと書いたが，それは，CALSのデータベースはCALS用のDTDに従って書かれていることを示す．

以前，「SGML方式」という言い方をしたが，その理由は，用語"SGML"の意味はDTDを書くための言語に限定し，DTDに基づいたDBやシステムについては「SGML方式」として区別するためであった．

なお，markupはmark upから作られた造語である．mark up（マーク付け）とは，編集者がゴチ，10ポなどの活字の指定やセンタリングや図表の位置などの割付け指示を原稿に赤字で入れることである．この指示を電子的に与えることをmarkupと一語で言う．マーク付けには固有と汎用がある．固有マーク付けの典型例は電算写植の印刷用制御コードで，これは使用するハード・ソフト固有である．一方，汎用の方は文書の体裁ではなく，構造を記述するmarkupで，文書の構成要素の始点と終点をマークする．こうすれば，マーク付け自体はハード・ソフトによらなくなるから汎用となる．

2.2 柔 軟 性

SGMLは柔軟性が高い．この性質はSGMLを他の規格と一緒に使うときに便利である．これもCALSにSGMLが採用された理由の1つだろう．

SGML方式ではテキスト・データだけでなく，図や写真なども扱えるようにしている．図や写真はビットパターン化したファイルとし，それを外部データ（SGMLの用語では外部エンティティという）として付ける．そして，それを文書のテキストから参照する．画像の形式にはGIF（Graphics Interchange Format），TIFF（Tag Image File Format），PS（PostScript）等々いろいろな種類があるが，SGML方式ではDTDで定義すればどれでも使える．

文字データについてもSGML方式は柔軟である．たとえば，数式や表はSGMLの書法でも書けるが，DTDで定義しておけば数式や表の部分のみLaTeXで書くこともできる．また，外字もDTDで名前を定義することによって，テキスト中では＆名前；の形式で書ける．名前はASCII文字で書くから，データの交換流通も安心してで

きる．

映像や音声についても外部エンティティとして扱うことができる．MPEG，QuickTimeなどの形式は問わない．ただし，これはファイルとして扱うので，その中身を分けて扱うわけではない．そこで，映像・音声の中身を扱うためのISO，JIS規格として，SGMLの拡張であるHyTime[13]が定められている．

2.3 SGML方式文書DBの構成

SGML宣言（SGML declaration），DTD，文書のテキスト，それに図や写真をビットパターン化したファイル（外部エンティティ）の4つで構成される．

SGML宣言には，使用する文字セットや処理系に要求するSGMLの機能などを書く．文書の言語が，英語，日本語など何で書かれていても，文字セットの指定で対応できる．

DTDは文書を構成する要素（element）とその相互関係，すなわち階層関係と参照関係をELEMENT文，ATTLIST文，ENTITY文を使って定義する．**図-1**に，単行本のDTDを例に挙げて説明する．

本は，タイトル，著者，目次などの前付け（front matter，以下fmと略），章・節・段落などで構成される本体（body，bdy），索引，奥書などの後付け（back matter，bm）といった階層関係を持っている．図-1の2行目で，element名"monogrf"

```
<!ENTITY %floats "fig | tabl | note">
<!ELEMENT monogrf - - (fm, bdy, bm?) +(%floats;)>
<!ELEMENT fm    - O (tit, sbt?, aut, toc)>
<!ELEMENT bdy   - O (chp+)>
<!ELEMENT bm    - O (indx?, pbls?, pbdt?)>
<!ELEMENT chp   - O (chptit, (sec+ | p+)>
<!ELEMENT sec   - O (sectit, p+)>
<!ELEMENT p     - O (#PCDATA)+>
   (中略)
<!ELEMENT fig   - - (nf?, figcap?, figbdy)>
<!ATTLIST fig   id  ID  #IMPLIED
          file CDATA #IMPLIED
          type CDATA #IMPLIED (中略)>
<!ELEMENT figref - O  EMPTY>
<!ATTLIST figref refid IDREF #IMPLIED>
   (以下略)
```

階層構造はELEMENT文を使って，図表などの参照関係はATTLIST文で定義する属性を使って表現する．

図-1 単行本のDTD（一部）

はfm，bdy，bmから構成され，出現順はfm，bdy，bmの順であることを示している．要素が順に出現するときはカンマで区切り，どちらかが現れるときは|（OR）を，両方が現れるが順は問わないときは&を使う．要素の出現頻度が，0または1のときは?で，1からN回のときは+で，0からN回のときは*で示す．たとえば，4行目は，本体は章の繰返しで構成されることを，6行目は，章は章のタイトルと，複数の節または段落で構成されることを表現している．なお，#PCDATAは解析対象文字データという意味で，予約語である．

ELEMENT文は要素の始点・終点のマークが省略できるか否かも示す．"O"は省略可，"-"は否を示す．左が始点で，右が終点を示す．

図表や注などは本文からの参照関係を持つ．そこで，まず1行目と2行目で，図，表または注などを%floatsとし，monogrfのどこにも現れうることを示す．そして，参照・被参照の関係をIDを使って表現する．IDは11行目や15行目のように属性リストの1つとして定義する．また，ファイル名は属性fileで，GIFほかの画像の形式は属性typeで示すことを定義すればよい（12，13行目参照）．

仮に「情報システム基盤技術としてのSGML」というタイトルの単行本があり，それが図-1のDTDに従うとすれば，その文書のテキストは**図-2**に示すようになる．

そして，図を参照する場合は，たとえば，図-2の下から3行目に示すように本文中には<figref refid=fig1>と書く．この記述はfig1をIDとして図を参照することを示している．図そのものは別ファイルにあるが，IDを使って参照するわけである．なお，図-1の下から3行目のfigrefのELEMENT文でEMPTYとあるのは，文書テキスト中にはfigrefのテキストデータがないことを示していたのである．表や注の場合も同じようにIDを使って参照できる．

ところで，このような<要素名>や</要素名>付きの複雑な文書テキストを書くのは面倒である．そこで，通常の文章を書く感覚で入力すれば，自動的にSGML方式の文書テキストに変換してくれるツールが開発されている．欧米だけでなく，最近は日本でもそれが普及してきたから，容易にSGML方式の文書テキストが作れるようになって

```
<monogrf>
<fm>
  <tit> 情報システム基盤技術としてのSGML
 <aut> 石塚英弘，根岸正光
  <toc>1. はじめに
       2. SGMLの趣旨
       (中略)
<bdy>
  <chp><chptit> はじめに
   <p> 欧米に比べて....だろうか.
   <p>SGMLは様々な情報システムで.....する.
          (中略)
  <chp><chptit>SGMLの趣旨
  <sec><sectit> 文書構造の表現
     <p>SGMLの趣旨は文書の構造を....有効である.
          (中略)
  <sec><sectit>SGML方式文書DBの構成
   <p>SGML宣言....の4つで構成される.
          (中略)
   <p>DTDは....図1<figref refid=fig1>に.....
    (以下大幅に略す)
</monogrf>
```

図-1のDTDに従って書いたSGML方式文書のテキスト．要素ごとに，その開始点を示すタグ＜要素名＞が付いている．

図-2 SGML方式文書のテキストの例

きた．また，SGML方式の文書テキストはSGML対応のDTPあるいはviewerを使えば綺麗にレイアウトされる[11,19]から，SGML方式の文書テキストはダンプ・リストのようなものと思っていただければよい．

また，SGML方式文書DBにはDTDが付いているから，それを基に別の構造の文書DBにツールを使って変換することもできる．そのため，SGML方式文書DBは交換流通に適しているのである．

3. HTMLの趣旨とSGMLとの相違点

HTMLはHyperText Markup Languageの略だが，これは文字どおりHTMLの趣旨を表している．すなわち，WWWというハイパテキスト・システム専用のDTDに従ったマーク付け記法によるデータ入力書式がHTMLである．HTMLは，WWWのハイパテキスト機能を実現するための書式になっているが，他の検索表示システムには対応しない．また，HTMLはWWWで表示するための記法であって，SGMLのように文書の構造を汎用的に表現するものではない．たとえば，章，節といった構造はない．また，見出し項目はあるが，著者という項目はない．

SGMLならば，DTDを自分で書くことができるが，HTMLの場合は書いたり修正することはできない．HTMLはWWWの機能に結びついた記法であり，もしもHTMLの記法を変えるとなるとWWWの機能を変えなければならないからである．

もっとも，SGML方式でDTDに従って書く文書テキストの記法はマーク付けであり，その点はHTMLによく似ている．そのため，「HTMLとSGMLは同じようなもんだ」と誤解する人や「HTMLはSGMLのサブセット」という誤解を呼ぶ言い方が出たのだろう．しかし，これまで説明したように，HTMLとSGMLは趣旨が異なるのだから，サブセットという言い方は適切でない．

WWWは確かに優れた検索表示ソフトウェアである．しかし，Mosaic, Netscape, HotJava[14]と少しづつ異なるソフトがあるし，もっと優れたソフトが次々に出てくる可能性が高い．表示に重点を置いたHTML形式よりは，構造を表現したSGML方式でデータを保持した方がソフトの変化に対応しやすいだろう．

なお，WWWサーバの中には，見た目はWWWでもデータはHTMLではなくSGML方式で持っているものがある．たとえば，データはSGML対応の検索エンジン，たとえばOpenTextの中に格納されており，検索はエンジンで実行し，結果を標示する時にHTMLに変換してクライアントに送るものがある．また最近では，Web上でSGML方式の文書DBを表示するviewerたとえばSoftQuad社のPanoramaもある．

4. ODA

ODA[15]-[17]は，1）表示用の割付け構造の交換を重視し，データベース的な構造（論理構造）だけでなく，表示用の割付け構造も文書構造として規格の対象としたこと，2）特定構造と共通構造があること．前者は利用者が読むことができる構造で，後者は文書の生成を導くテンプレートである．3）文書構造モデルにオブジェクト指向のクラスの概念を取り入れたこと，の3点でSGMLと異なっている．

ODAでは，論理構造と割付け構造は別個のもので，原則として無関係としている．ただし，割

付け処理は，論理構造と結びついている割付け指示によって制御されることもあるとしている．一方，SGML方式では割付け構造は別の規格：DSSSL（Document Style Semantics and Specification Language）やSPDL（Standard Page Description Language）とし，SGMLからDSSSL，DSSSLからSPDLへの変換機能が規定されている．

ODAでは，文書の構成要素を対象体（object）というが，SGMLと同様，階層関係と参照関係，順序，繰返しなどの機能をサポートする．

ODAで規定した文書の表現形式は，1）書式付き形式：この形式では，発信者が意図したとおりに文書の表示ができる，2）処理可能形式：この形式では，文書の編集および書式付けができる，3）書式付き処理可能形式: 文書の表示・編集・書式付けができる，の3つである．

ODAの文書交換様式ODIFには，クラスAと同Bとがある．Bは，書式付き文書の交換に用いる．一方，Aは論理構造・割付け構造どちらも表現できる．なお，ODIFではSGML文書交換様式（SDIF）にも対応するため，交換用の言語ODL（事務文書用言語）も付属書で規定している．ODLはSGMLのアプリケーションの1つである．

5. 情報システム基盤としてのSGML方式文書DB

文書データは関係型DBMSによる従来の情報システムでは扱いにくい対象であった．しかし，SGML方式文書DBとすることで，たとえばCALSに見られるように，情報システムの中で，ネットワークの中で，他のデータと共に有機的に操作できる対象となった．

そして，SGML方式文書DBの中から，ある業務に必要なデータのみを取り出し，その業務に適したレイアウトで見せたり，印刷したり，ネットワークで送ったりすることもできるようになった．これは新形態の電子出版と観ることもできる．なぜなら，電子出版とは出版の各工程をコンピュータ関連技術で行うことであり，旧来の編集，版下作成，印刷，製本，配布から，編集，ネットワークを通しての送受信，検索，表示に変わってきたからである．

SGML方式文書DBはCALSのスタートがそうであったように，当初，マニュアルの世界で実用化した．その後，様々な分野で実用化している．学術情報の分野でも欧米の大手出版社がSGML方式文書DBに基づく出版[18]に乗り出している．また日本でも，（国立）学術情報センターの取組み[12]がある．また，日本化学会は93年1月号から欧文論文誌のSGML方式による電子出版を開始し，毎月約60論文から成る論文誌を印刷している．そして，WWWによる実験サービスやCD-ROMの試作も行い，SGML方式文書DBから様々な出版ができることを実証した[19]．さらに，SGML方式文書DBの対象分野は理工系に留まらず，人文科学分野，たとえば古典文学のテキスト・データベースにも広がっている[20]．

この種の学術情報のSGML方式文書DBがネットワークを通して提供されるようになると，電子図書館[21]が現実味を帯びてくる．机の上に置いたWSあるいはPCから世界の図書館の文献を検索表示し，それを読みながら調査・研究・実験し，レポートや論文をWSやPCで作成してネットワークに投稿する．近い将来，それが普通になるかもしれない．

6. CALSの一例：ソフトウェアCALS

我が国でもNCALSほか，CALSへの取組みが進んできた．本稿ではCALS-SGMLの適用性の広さを示す一例として，95年11月にスタートしたソフトウェアCALSのプロジェクトを簡単に紹介する．

名称はソフトウェアCALSに関する調査研究委員会，委員長は相磯慶應大学大学院教授，通産省工業技術院，通産省，NTT，コンピュータ・メーカ系企業，ユーザ系企業，ソフトハウス系企業，学識経験者などが参加し，事務局は日本規格協会である．

このプロジェクトはソフトウェア開発をいかに効率化するかをCALSの観点から検討しようとするものである．ごく近い将来，ネットワーク環境下で複数企業が協同してソフトウェアを開発するようになるとの認識の下に，仕様書ほかのドキュメントやCASEデータなどのプロダクト・データの交換形式，管理データの交換形式，データ・セキュリティなどの問題を調査研究する．始まったばかりであるが，WWW (http://goran.sl.cae.ntt.jp/) を使って情報を公開しようとしていること，実質

的に役に立つことを目指している点で今後が注目される．

7. おわりに代えて

CALSの浸透にともない，SGMLは日本でも常識化するだろう．すなわち，SGMLツールが普及し，SGMLの面倒な点は意識せずに使えるようになろう．事実，最近のツールの普及は著しい．近い将来，情報システムがSGML方式文書DBを扱うことが一般化するだろう．そして，エンドユーザはSGMLを意識せずに自然に使うことになろう．

参 考 文 献

1) ISO 8879, Information processing - Text and office systems - Standard Generalied Markup Language (SGML).
2) JIS X 4151, 文書記述言語 SGML, 日本規格協会.
3) 後藤龍男：CALS: 21世紀における企業情報システムの国際標準確立と企業統合に向けて，情報処理, Vol.36, No.1, pp.1-7 (1995).
4) 水田 浩：CALSの可能性, 256p., 生産性出版(1995).
5) たとえば，益岡竜介，木庭袋圭裕：World-Wide Web, 情報処理, Vol.36, No.12, pp.1155-1165 (1995).
6) 吉岡 誠編著：SGMLのススメ, 167p., オーム社 (1993).
7) 根岸正光，石塚英弘共編：SGMLの活用, 168p., オーム社 (1994).
8) Eric van Herwijinen 著, SGML 懇談会実用化WG監訳：実践SGML, 日本規格協会 (1992).
9) Martin Bryan 著, 福島 誠訳：SGML入門, アスキー出版局 (1991).
10) Goldfarb, C.F. : The SGML Handbook, 663p., Oxford University Press (1990).
11) 田中洋一：文書記述言語SGMLとその動向, 情報処理, Vol.32, No.10, pp.1118-1125 (1991).
12)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, Vol.33, No.4, pp.413-420 (1992).
13) 小町祐史：マルチメディア／ハイパメディア情報交換の標準化動向, 情報処理, Vol.35, No.7, pp.632-641 (1994); 小町祐史：HyTimeの規定，文献7)のpp.90-99 (1994).
14) Gosling, J. and McGilton, H. : The Java Language Environment A White Paper, 65p. Sun Microsystems, (May 1995) (Sun Microsystemsによる翻訳あり).
15)若島睦夫，坂入 隆，真野芳久：ODA：多様な文書のための標準様式, 情報処理, Vol.37, No.3, pp.199-206 (1996).
16) ISO 8613, Information Processing - Text and Office Systems - Office Document Architecture (ODA) and Interchange Format.
17) JIS X 4101, 開放型文書体系(ODA)及び交換様式－第1部 総則; JIS X 4102, 同－第2部 文書構造; JIS X 4104, 同－第4部 文書概要; JIS X 4105, 同－第5部 開放型文書交換様式(ODIF); JIS X 4106, 同－第6部 文字内容体系; JIS X 4107, 同－第7部 ラスタ図形内容体系; JIS X 4108, 同－第8部 幾何学図形内容体系.
18) 根岸正光：SGML普及への展望，文献7)のpp.144-164 (1994).
19) 石塚英弘：電子出版 その概念と技術, 電子情報通信学会誌，Vol.78, No.9, pp.891-898(1995); H.Ishizuka: The Reception of SGML Based Electronic Publishing by Japanese Scientific Community, Proc. 47th FID (Int. Fed. Inf. Doc.) Conf. Cong., pp.505-508 (Oct. 1994, Omiya).
20) 長瀬真理：テキスト・データベースとTEI, 文献7)のpp.117-141 (1994).
21) 長尾 真：電子図書館, 岩波書店 (1994).

（平成8年1月11日受付）

石塚 英弘（正会員）

1946年生．1969年東京大学理学部化学科卒業．1974年同大学院理学系研究科博士課程修了，理学博士．同年，東京大学助手．1976年国文学研究資料館研究情報部助教授．1982年図書館情報大学助教授を経て，1992年同大教授．全文データベース・電子図書システム，電子図書館，情報知識システム等に興味を持つ．電子情報通信学会，ACM等各会員．

根岸 正光（正会員）

1945年生．1968年東京大学経済学部卒業．1976年同大学院経済学研究科博士課程修了．東京大学助手，講師，助教授（情報図書館学研究センター）を経て，1986年学術情報センター教授（データベース研究部門），1994年同センター研究主幹（研究動向調査研究系）．オンライン情報検索システム，学術雑誌総合目録データベース，全文データベース・システム等の研究開発等に従事．最近は，電子図書館，計量書誌学的方法による研究水準の国際比較等を研究．ASIS会員．