<情報知識学会SGML研修フォーラム 1997.2.5>

[総論] SGMLと電子文書の現況と将来

根岸正光*
文部省学術情報センター

1. はじめに

筆者がSGMLなることを初めて耳にしたのは、1987年10月に米国のCAS (Chemical Abstracts Service)を訪れたときであった。[1] 当時われわれは全文（フルテキスト）データベースのための新たなシステム開発に向けて、調査研究を進めており、その一環として、この分野の先達ともいえるCASを訪ね、これに関わる技術や体制などについて、いわば教えを請うといった塩梅であった。爾来すでにまる10年になろうとするところであるが、わが国におけるSGMLの応用は、あまりはかばかしいようには見えないが、はたしてそうか。[2,3]

現下、いわゆるマルチメディア・ブームのさ中、電子化文書、電子出版、あるいは電子図書館等々が語られる状況において、SGMLを中心として、HTMLなど関連する規格を含め、これらの将来を検討してみたい。

2. 電子文書の規格

2.1 SGMLとフォーマッター

SGMLは、いうまでもなく、文書を電子する際に適用されるべき規格のひとつである。筆者らがSGMLについて検討を始めた当時、すなわち1990年前後の時点では、一方においてすでに特定の利用者層にTeXが浸透しており、片やOSIの一環として、ODAが推進されつつあった。そこで、当時の議論は、これら三者の優劣を比較衡量するという体裁となった。この場合、SGMLとODAは一応規格が制定されたところであるが、現実にこれを応用するためのソフトウエア製品は未開発の段階であった。一方、TeXは広汎ではないとしても、すでに実用されていて、そのいみでは説得力のあるものとなっていた。

さらに、TeXは印刷物を電子的に作成する道具、すなわちフォーマッターであり、従ってその効用も即物的でわかりやすい。一方、SGMLやODAは電子文書の交換・流通を目的とし、抽象化、論理化された文書を対象にするところから、まず以てわかりにくさがつきまとう点は否定できない。これは、そもそも電子文書の電子的交換・流通ということ自体がなく、せいぜいワープロ文書ファイルの交換が、機種依存性にまつわる多くの難点を伴いながらも、ある程度行われていたに過ぎないという当時の状況からみれば、致し方のないことでもあったといえよう。

*Masamitsu NEGISHI (NACSIS: National Cener for Science Information Systems)

2.2 紙媒体モデルと電子文書モデル

紙媒体モデル —— 「文書」といえば、今日でも常識的には紙媒体の上に定着されるものという理解が普通であろう。従って、これを電子化して交換・流通させるといっても、最終産物はあくまで印刷物なのであって、その流通の中間段階に電子的手段を用いて、なにがしかの効率化を図るとしても、それは結局は印刷工程の電算化の範囲内に止まることになる。そして、交換・流通における重大な要件は、印刷物としての版面の厳密な再現性となる。つまり、そこでは印刷機、複写機、Faxの機能と同等のものが求められており、電子化によって、こうした機能がいかほど安上がりになるのかが、ほとんど唯一の関心事になるのである。こうした文脈では、たとえ規格が制定されたといっても、それはいわば印刷技術に関わる問題と解され、利用者、編集者の直接関知するところとはなるまい。

自家発行文化 —— このような状況に劇的変化をもたらしたのが、インターネットである。そこでは、まず文書の作成から流通までの全体が、著者本人の仕事とみなされるようになった。これはすなわち大衆参加による自家発行の文化といってもよい。ここで「発行」とは、現況ではWWWによるものが主に想定されるが、電子メールやニューズ・グループなど、WWW出現以前からのものも含めて考えたい。これらによって、電子的な文書の交換が身近な現実となったからである。ところで、こうした原初的段階では、単純なASCII平文のみが対象とされたが、それでも日本語に関しては、文字コードの問題があって、その扱いは現今でもそれなりに厄介ではある。

電子文書モデル／ハイパーテキスト —— WWWの出現は、文書について、従来の紙媒体モデルから離れて、リンクを使って関連箇所を自在に飛び回るという点を特徴とする、電子文書モデルを現実に提示して見せた。それ以前にも、主としてCD-ROMを用いて、リンク結合を実現し、また画像・音声を貼り込んでマルチメディア化することは行われてきたが、当時の感覚として、その大容量に着目し、辞書、事典、名鑑類をCD-ROM化するという、いわばデータベース的応用が主体となった。そして、わが国では、この種の内容に主眼をおいて、EPWINGなる規格も作られた。

このほか、リンク結合を特徴とする電子文書のためのシステムとしては、Macintoshのハイパーカードが先行したが、その普及は当時の高級機であるMacintoshの普及度合いに限定されたものであった。結局、ハイパーテキストの本格的普及はWWWなるインターネット上のアプリケーションによってはじめて実現したことになる。

3. HTMLの動向

3.1 HTMLとSGML

WWW文書の記述規格はHTMLであるが、これとSGMLとの関係は、発祥的にはHTMLはSGMLのサブセット、あるいは応用規格であるということになるであろう。すなわち、HTMLは既定の単一DTDを前提とした文書記述方式であるからである。そして、これはそのためのフォーマッター、すなわちWWWでいうブラウザーに直結しているという体裁である。すなわち、HTMLは表示画面上あるいは印刷紙面での表現形式と直結しており、SGMLにおける文書構造の論理表現といった抽象的なわかりにくさはない。この点で、HTMLは利用者からみれば、あたかもTeXと同様のものと認識され

る。

もっとも、実際の表示画面では、ブラウザーによって差異が生じる。これは、文書の論理構造と具体的媒体上の表現の関係という、SGMLにとって基本的に重大な要素の現れなのであるが、一般には、こうした深刻な認識などは到底なく、むしろブラウザーの優劣という方向に議論が走ってしまう。これは無理からぬところとはいえ、われわれにとって興味深いものがある。

3.2 WWW文書の魅力

SGMLを適用する際の大きな問題として、タグの挿入の工数がある。すなわち、文書にタグを入れて行くのは大変な手間であり、到底一般の著者などができるものではなく、この点で、余程充実したオーサリング・ソフトが出回らない限り、SGMLを使うわけにはゆかないといった主張が支配的であった。HTMLの場合も、いかにタグの種類が少ないとはいえ、まさにSGML方式によるタグ入れ作業が必要で、従来の議論からすれば、工数的にみて、これは到底容認できるようなものでない。従って、HTMLについても、実用性がないということになるはずであった。

情報発信機構 —— しかし、実際には、多くの者がむしろ嬉々としてHTML文書づくりにいそしむという光景が現出されたのは、どうしてか。これは、そのようにして作成されるHTML文書が、たとえ擬制的にもせよ、即座に世界に向けて発信されるという、これまで想像もできなかったようなシステム、体制が実現したからである。こうした目に見える効果・効用の前で、著者たちは、てまひまをいとわず、面倒なHTMLのタグ入れに励んだものと思われる。

そして程なく、タグ入れを支援するさまざまなツール・ソフトが、これまた関係の有志によって開発され、PDSとして配布された。こうした状況のなかで、ワープロ・ソフト製品開発者の方でも、HTML作成支援機能を導入付加することが進められ、今や主要なパソコン・ワープロ・ソフト製品では、すべてHTML文書の作成機能を持つようになっている。そこでは、関連のプロバイダーのWebサイトへのアップロードまでを、一貫してできるような工夫も施され、ホームページづくりの大衆化が促されている。こうしたソフト開発者側としては、インターネットを用いたビジネスの展開を企図しており、ホームページづくりの普及もその重要な要素である。そこで、ワープロの機能を向上させて、その販売を促進するのと、プロバイダー関連事業の推進との二股かけた経営戦略として、これは至極もっともなりゆきである。

マルチメディア文書 —— HTML普及の要素としては、上述のような情報発信機構との連動のほか、表現方法における面白さがあると考えられる。すなわち、すでにふれたとおり、HTMLでは、従来の文書の概念を超えた電子文書が意図されているからである。具体的には、リンク結合によるハイパーテキスト化と、画像、音声等を貼り込むマルチメディア化である。HTMLを用いることにより、従来の紙の文書では不可能な、面白いページづくりができる。このため、個人レベルではホームページ作成という新しい趣味の世界が開かれた。一方、企業的応用においては、HTML文書の顧客に対する訴求力、吸引力が注目されて、WWWは今や重要な宣伝媒体と位置付けられるようになっており、この方面でもHTMLの応用は普及している。こうして、HTMLは個人からビジネス利用まで、幅広い支持を得た格好である。

3.3 HTMLの独自拡張と空洞化

ところで、HTMLに関わる最近の状況の展開は、我々に興味ある点を提示している。すなわち、個々のブラウザー作成者によって、規格が独自に拡張される一方、プラグ・インを用いた表現が多用されるようになって、HTML自体は実質的に空洞化してきているという点である。HTML規約の独自の拡張は、ブラウザー市場の囲い込みを企図して行われたものであるが、プラグ・イン機構の導入とともに、目新しい表現方法をWeb上で実現するべく、マルチメディア系ソフトの各社がその製品のプラグ・イン対応を図って、ここに新たな市場を求めつつある。こうして、HTMLは今やこれらプラグ・インの単なる呼出し機構的性格が強くなっている。しかし、この点をむしろ前向きに、WWWのOS化であると捉えて、WWWが、今後コンピュータの標準ユーザー・インターフェースになるという考え方も現れ、事実この方向での開発も進みつつあるようである。まさに、事態は当初思いもよらない展開を見せつつあるといってよかろう。

3.4 PDFの動向

最近になって、わが国でも、電子文書関連のあらたな規格・方式として、Adobe社のPDFが注目されるようになってきた。PDFは数年前から米国で普及してきたが、昨年、同社がその日本語対応を発表するに及んで、わが国でもにわかに注目されるようになったものである。PDFは、そのビューアーを同社が無料で配布したことが普及の原動力になっている。ともあれ、PDFはPostScriptのネットワーク対応版といった体裁であり、従ってSGMLとはレベルを異にするものであるが、しかしその分、TeXと同様に版面の再生用というわかりやすさがある。さらに、仮想的プリンターに出力するという体裁で、PDF文書が安直に生成できることから、その普及を予想する向きもある。もっとも、日本語文書の場合、フォントの処理において、ファイル・サイズの増大という問題ばかりでなく、フォントの著作権確保の問題も生じるので、なりゆきが注目されるところである。

4. SGMLの現況

4.1 SGMLの応用分野

ここで、SGMLの現況について、主として米国の状況をみてみよう。SGML関連の有力な情報源となっている米国のThe Summer Institute of LinguisticsのWebページには、SGMLの紹介から始まって、関連規格、文献案内、協議会等関連団体、ソフトウエア、適用事業など、一通りの情報が網羅されていて非常に便利である。[4] そこには、現時点におけるSGMLの一般的な適用活動として、下記の37件が掲げられている。これらについて個々に紹介することはできないが、SGMLを巡って、関連する規格の制定や適用事業が多方面にわたって進められていることが看取されるであろう。

(1) HyTime: ISO 10744 Hypermedia/Time-based Structuring Language
(2) SMDL - Standard Music Description Language, ISO/IEC DIS 10743:1995
(3) SGML Initiative in Health Care (HL7 Health Level-7 and SGML)
(4) Metafile for Interactive Documents (MID)

(5) Standard Hypermedia/Multimedia Scripting Language (SMSL)
(6) Digital Libraries (Initiative) and SGML
(7) SGML and Metadata
(8) Hyper-G Text Format (HTF)
(9) Association of American Publishers (AAP)
(10) ISO 12083 DTDs
(11) IBMIDDoc: IBM Information Development document type
(12) IEEE Standards Department
(13) Davenport Group: DocBook DTD
(14) ICADD: International Committee on Accessible Document Design
(15) CAPS (Communication and Access to Information for Persons with Special Needs) and HARMONY (Horizontal Action for the Harmonisation of Accessible Structured Documents)
(16) ELVIS - Elektronisches Literaturverzeichnis - Informatik fur Sehgeschadigte
(17) NITF (News Industry Text Format) [Formerly UTF - Universal Text Format] - SGML for the News Distribution Industry
(18) Canadian Strategic Software Consortium (CSSC): SGML and SQL
(19) Electronic PROTEIN SCIENCE
(20) MIME-SGML (Multipurpose Internet Mail Extensions)
(21) EWS-MAJOUR
(22) OCLC SGML Projects
(23) SGML and Chemistry: The OCLC CORE Project (Chemistry Online Retrieval Experiment) and other Initiatives
(24) Chemical Markup Language (CLM)
(25) SGML and Chemistry - Other Links
(27) SGML and Physics: The American Physical Society, American Astronomical Society, and The American Institute of Physics
(28) GMD-IPSI SGML Projects
(29) Multiagency Electronic [Pharmaceutical] Regulatory Submission (MERS) Project
(30) Joint Electronic Document Interchange (JEDI)
(31) Earth Interactions: An Electronic Journal in SGML
(32) Topic Navigation Maps
(33) Information Mapping and SGML
(34) The Corpus Legis Project
(35) Royal Melbourne Institute of Technology (RMIT) - SIM SGML Database Technology
(36) EUROMATH Project
(37) SSML: A Speech Synthesis Markup Language

さらに、大学関係プロジェクトとしては、Text Encoding Initiative (TEI)、Oxford Text Archive (OTA)、University of Virginia Electronic Text Center、British National Corpus Project (BNC)、American Memory Project, Library of Congress 等々有名なものを含めて約70件が

掲載されている。また、政府および業界プロジェクトとして、The Department of Energy (DOE) Office of Scientific and Technical Information (OSTI) 、IRS (United States Internal Revenue Service) 、National Library of Medicine (NLM) 、Library of Congress - Encoded Archival Description (EAD) - Finding Aid Pilot Project など18件が掲げられている。CALS (Continuous Acquisition and Lifecycle Support)についても、ここに紹介され、関係サイトへのリンクが張られている。

これらをみると、SGMLは、Hytime、SMDLなど、マルチメディア系の内容への応用ないし規格の拡張が模索される一方、実際の適用場面では、当然ながら、全文データベースや電子図書館など、電子文書の集積と配布についての応用が種々試みられていることがわかる。

4.2 Matadataへの応用 —— Dublin Core

また、これらとは別に注目するべきものとして、Metadataへの応用がある。SGMLは、元来電子文書それ自体のための規格ということを主眼に考案されたものと考えられるが、インターネットの普及に伴う、各種各様の電子文書の氾濫の中で、これら電子文書の探索・参照のためのツールとしても応用しようという発想が生まれている。

すなわち、OCLCとNCSAが主体となって進められているDublin Core計画である。[5] これは一口にいえば、各種電子文書の目録データベースを構成しようという試みである。従来、図書館で行われてきた目録作成を、インターネット上の電子化資料にも拡張して、効率的な資料探索ができるようにしようとするものである。すでにLycos、AltaVista、Open Text Index等々、いわゆるサーチ・エンジンが運用され、よく用いられているが、こうした機械的索引方式では、検索がきわめて不効率であることが実感されている。一方、従来の図書館目録と同様に、目録専門家が、膨大なネットワーク情報資源を逐一点検して、目録データを作成してゆくことは到底不可能である。

ここにおいて、自らの電子文書が有効に検索されることを望む著者が、一定規格の目録データを各自作成して、文書に付加することにし、この規格に対応するサーチ・エンジンでこれらを探索するようにすれば、事態は随分改善されるであろう。つまりこれはMARCのインターネット版といもいえる。こうした考えに基づいて、Dublin Core計画では、著者、表題、主題等、必要最小限の記述項目を設定し、そのSGMLによる記述規格(DTD)を設定しようという提案を行っている。現時点で、その成否はもちろん不明であるが、ネットワーク情報資源の有効活用という視点ではきわめて妥当な解決策と思われ、その進展に期待したい。

4.3 SGML関連ソフトウエアの現況

次にSGML関連ソフトウエアの開発状況をみてみると、同じくこのページに、Public Softwareとして、Parser、Editor, Browser, Converter, Formatter, DSSSL toolなどの40件ほどが紹介されている。また、商用ソフト製品に関しては、NICE technologies社が調査を行っており、その結果として、90件ほどのソフトウエアが同社のWWWで公開されている。[6] そこにはさらに、Editors: 16件、Convertors: 19、Browsers: 12、Databases: 16、Page layout tools: 5、Web interfaces: 2、Document analysis tools: 4、DTD design tools: 3、Parsers: 6、HyTime systems: 1、合計84件といった集計も出ていて興味深い。その解説によれば、Conversion toolが多くを占めるのは、SGMLがデータ作成・入力の形式とし

てはそれほど使われておらず、他の方法で作成されたデータをSGML形式で保存・蓄積するという応用が多いことの現れであろうとしている。従って、それらを配信するためのBrowserとかWeb interfaceも多くなっている。

5. 学術情報センターのSGML全文データベースと『電子図書館』

学術情報センターではSGMLの能力や将来性に以前から着目し、学会誌をSGMLで編集する実験をし、学会にもSGMLでの学会誌編集を呼びかけてきた。その結果、SGMLを用いて学会誌を編集する学会も現れてきている。

センターでは、現在、学会誌のページ・イメージを集積する方式による「電子図書館システム」を開発し、1997年4月から公開サービスに供することにしているが、その一方で、従来から行っているデータベース・サービスについては、全文検索エンジンを利用したあらたな検索システムに全面的に移行しつつある。そこでは、全文データベースのみならず、文献抄録型や引用索引型のデータベースを含めて、すべてのデータベースが内部的にはSGMLによって記述されることになる。これにより、レコード内のデータ構造、すなわち全文データベースの場合には文書の論理構造に即したこまかな検索も可能になる。[7)]

この検索システムでは、従来風のコマンド型インターフェースの他、Webブラウザーによる GUIでの検索システムも開発、実験中である[http://www. rd. nacsis. ac. jp:8080 /PatMosaic/]。この場合、データは当然HTMLで送る必要があり、データベースの内部的SGML形式からHTMLへ動的にデータを変換して送信するようになっている。また必要に応じて、SGML形式でデータを受けとることもできる。この方法により、検索機能が向上するだけでなく、表示においても、例えば論文の目次を表示するとか、ある号の雑誌の著者一覧を見るなど、多様な指定が可能になる。

前記の「電子図書館システム」では、当面既存の学会誌を効率的に収容してゆくため、版面イメージ主体のデータベースを構成することににしているが、学会誌編集における電子化、なかんずくSGML適用の進展に従って、全文データベース・システムとの融合が図られ、学術情報の効率的流通が一層促進されることが期待される。

6. おわりに —— プロトコル／ソフトウエア規格の特性

これまで、各種の工業規格が制定され、これが産業の発達に大いに寄与してきたことは事実である。これら従来の規格は、材料の組成や機器の寸法など、化学的、物理的あるいは工学的な合理性を根拠としている。

これに対して、本稿の主題であるSGMLをはじめとする、プロトコルあるいはソフトウエア関連の規格の類は、だいぶ様子が異なるようにみえる。元来プロトコルは約束事、決めごとであって、化学・物理学などの自然科学的法則からくる制約は少ない。もっとも、目的合理性に強く制約されることから、設計上さほどの自由度があるわけではなく、実体的な差別化は難しい。しかし、プロトコルは些細な食い違いでも相互接続できないという非寛容的性格があり、この点をビジネス・チャンスとして、いかに自己に有利に利用するかという視点で、熾烈な競争が展開されているというのが、ソフトウエア業界の現状であろう。

ここでは、規格は市場囲い込みの手段であり、「デファクト・スタンダード」の地位獲得を巡ってさまざまな戦略が採られるが、その際の有力な方策として、ソフトウエアの無料配布がある。これは強大な資金力を背景として可能になるもので、資金的にこうした政策に追随できない競争者を市場から追放して、独占的地位を確保できる。このような粗野な市場支配も行われる中にあって、SGML関連は随分地味な存在にみえるが、これは、SGMLがオープンな規格として確立されているからであろう。そこでは、自由市場の本来に則して、利用者側における選択が真に重要な意味をもつといえる。この際、わが国においても、着実な応用を推進するとともに、利用者間での緊密な情報交換を図ってゆく必要があるであろう。

＜参考＞

1) 根岸正光、石塚英弘共編「ＳＧＭＬの活用」、オーム社、1994.12、168p. ISBN:4-274-07808-6.

2) 石塚英弘、根岸正光「情報システムの基盤技術としてのSGML — 文書データベースからWWWそしてCALSまで」、情報処理、Vol.37, No.3, p.207-212 (1996).

3) 根岸正光「電子文書・電子出版から電子取引まで — SGMLをめぐる諸活動」、第10回『大学と科学』公開シンポジウム組織委員会編「情報スーパーハイウエイ — 加速する研究・教育・医療」、クバプロ、1996、p.66-79. ISBN:4-906347-57-6

4) Robin Cover, "The SGML Web Page," [http://www.sil.org/sgml/]

5) Stuart Weibel, et al., "OCLC/NCSA Metadata Workshop Report," [http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html]

6) "NICE Technologies SGML Product List," [http://www.nicetech.com/venprod.htm]

7) 大山敬三「インターネットに適応した全文データベース検索システムの構成」学術情報センター紀要、第7号(1995). [http://www.rd.nacsis.ac.jp/~oyama/paper/kiyo-95/paper.html]