SGMLとインターネット／イントラネットとの連携

小林憲夫
norio@seishosha.co.jp
青松社　企画制作部

SGMLはISO 8879で制定された、古くて新しい文書標準化の規約である。CALSのコアテクノロジーとしてそのメリットは十分に認められながらも、これまで日本ではその使いにくさから、SGMLの普及はいっこうに進展していなかった。ところが昨今のインターネットの急速な浸透、そしてイントラネットの拡大が、SGMLの再認識をもたらすのではないかという観測が生じてきている。ここではSGMLが企業にもたらすインパクトを、最新の情報システムの動向と絡めながら、いかに有効に利用すればよいかという立場から述べている。

COLLABORATION BETWEEN SGML AND THE INTERNET/INTRANET

Norio Kobayashi

Planning and Production Center, Seishosha Co.,Ltd.

SGML is regulated by ISO 8879 in 1986, and a document standardizing method which is still prevailing. Although an advantage of SGML is widely recognized as a core technology for CALS, SGML has not been applied to so many fields mainly because of a difficulty of its handling. Recently, a new hope that a big movement and expansion regarding Internet and Intranet might cause re-evaluation of SGML technology is appeared. This paper introduces what kind and how much impact is happened to an industrial use of SGML with a newest information technology trend including RDBMS from a viewpoint of an effective use.

SGML: Standard Generalized Markup Language
CALS: Continuous Acquisition and Life-cycle Support
RDBMS: Relational Data Base Management System

1 はじめに

SGMLは、その淵源を1960年代末のIBM研究所にまでさかのぼることのできる文書標準化規約である。この規約は、1986年にISO 8879で制定されて以降、単に一企業だけの技術ではなく、世界的な規約として注目され普及が期待されてきた。世界中に存在する様々な文書の規約を統一することによって生じるメリットは、想像するだけでも明らかである。

さらにSGMLは、文書だけでなく様々なデータを標準化しようというCALSにも利用されている。むしろCALS内ではSGMLだけが標準化されているといっても過言ではない状況である。SGMLは、あらゆる意味で現代に不可欠のテクノロジーとして認められている。

しかし、その期待される効果とは逆に、現実にはSGMLの導入は必ずしも円滑に進行していない。タグ付けされたテキストは、DTPに慣れ親しんだ目からは異様に映る。構造化言語のために、レンダリング系を別個に設計しなければならない難しさもある。詳細に設計すればするほど莫大なプログラムを作成する必要がある。

こうした使いにくさから、日本ではSGMLの普及はいっこうに進展せず、わずかに政府主導の特定の分野や米国対応で利用されているにすぎない。ところが昨今のインターネットの急速な浸透、そしてイントラネットの拡大が事態を変えつつある。巨大なデータベースを情報リソースとして利用するWWWは、SGMLという標準文書形式とのマッチングが極めて勝れている。情報共有の形態としてのSGMLが、再認識されようとしているのだ。

本論では、最初にSGMLの概念を簡単に説明し、SGMLの問題点なども明確にした上で、インターネット／イントラネットがどの点でSGMLに有利かの説明を試みている。その後に、SGMLとインターネット／イントラネットとの関係が将来的にどうなるかを、現時点で当社が導入しつつある方法と共に紹介する。

ISO: International Standard Organization
SGML: Standard Generalized Markup Language
CALS: Continuous Acquisition and Life-cycle Support
DTP: Desk Top Publishing
WWW: World Wide Web

2 SGMLとはどんなものなのか

2.1 SGMLの基本原理と3つの要素

SGMLは3つの要素から構成される。SGML宣言とDTDそして文書インスタンスである。SGML宣言は利用するスタイル、ハードウェアやキャラクタセットなどの環境を記述したものであり、これによって利用条件が明確になる。

DTDはSGMLの核ともなるプログラム形式の記述ファイルである。利用するタグの定義をここで行なうと共に、タグで挟まれる文章の種類や形式についても規定する。通常のタグの種類は60から100種類であるが、表組みなどの情報も細かく定義すると200種類以上のタグとなり、数百行のDTDを作成しなければならない。

文書インスタンスは、SGMLのタグを付けた本文テキストのことである。ほとんどの場合、タグを含めてASCIIテキスト形式で書かれているため、プラットフォームに依存せずにデータのハンドリングが可能である。ただしインスタンスが日本語の場合、文字コードがマシンによって異なるために注意が必要である。

2.2 ISO8879による国際標準規約

SGMLは1960年代末に米国IBM社のCharles Goldfarb博士が社内文書の標準規約として開発した。その後社内の統一規約としてGMLの名称の元で1970年代末に一定の完成を見た。

ISOでは1980年以降標準化の検討を開始し、SGMLという名称と共に1986年に国際文書構造規約として認められた。その後ISO 9069（SDIF:SGMLドキュメント交換規

約)、ISO 9541（フォント規約）などが整備され、今日にいたっている。

2.3 構造化言語と組版言語

構造化言語と呼ばれる SGML が従来の組版言語と大きく異なるのは、文章を構造とレイアウトの二つに分解した点にある。すべての文章は構造を持ち、さらにそれが表現されるメディア（媒体）によって、同じ構造でも書体やレイアウトが異なるという特性を持つ。

SGML は、これまで混合して用いられていた構造とレイアウトとを分離することによって、異なるメディア間相互のデータ交換を可能にし、情報のハンドリングを飛躍的に高めることができる。プリンタのフォントの交換、用紙サイズの変更、CD-ROM での配布、データベースでの利用、データの流用、すべてにわたって柔軟に対応が可能となる。

DTD: Document Type Definition
GML: Generalized Markup Language
ASCII: American Standard Code for Information Interchange
SDIF: SGML Document Interchange Format
CD-ROM:Compact Disk Read Only Memory

3 青松社の SGML システム

3.1 青松社はなぜ SGML なのか

青松社は印刷会社であり、受注産業の例にもれず、システム全体として SGML を構築することはクライアントとの関係で不可能である。むしろマニュアルなどを SGML 化することにより、組版系を自動的に処理できるようにするのが導入の目的であった。

3.2 自動組版のための SGML

構造の解析や DTD 作成などにはコストが要求されるが、いったん作成したインスタンスは自動的に組版処理を行なえるので全体的なコストダウンを実現できる。特に翻訳ではレイアウトがすでに決まっていることと、それほど複雑なレイアウトでないので、SGML 導入に適していると言える。

3.3 全ドキュメントを SGML 化する

当社では受注した翻訳業務のうち、組版ソフトの指定がない大量のものに関してはほとんどの場合 SGML 化している。この経験により、DTD の効率的な作成技術と、レンダリング系とのマッチングによるトータルな品質向上のノウハウを蓄積できた。

4 SGML 導入におけるメリット

4.1 メディアへの対応柔軟性

当社にとって基本的な SGML 化によるメリットは、組版を自動化できること、他メディアへの変換の容易性と、さらにポータビリティ（絶対容量のコンパクト性）である。最後の点は処理系の負荷を減らすという面でも有効である。

システムとして SGML を構築している訳ではないので、データの再利用はほとんどない。しかし例えば、最近急速に注目されているインターネット向けの HTML は、固定 DTD による SGML サブセットであり、SGML データからの変換は容易である。

さらに SGML から PS ファイル化して PDF へ変換するニーズも増大している。紙出力も、組版マクロの工夫によりかなり高度な水準を達成している。そもそものきっかけは組版コストの削減と納期の短縮であったが、SGML 化することでさまざまな展開の可能性が開けた。

4.2 SGML をデータベースに利用する

マニュアルなどの書き起こしでは、SGML 化するために構造的な企画設計が必要となる。新規に書き起こす場合には、一般の SGML エディタと呼ばれるツールが有効である。しかしこの場合も、レイアウトを確認するためには組版マクロまでの一貫したシステムを用意しなければならない。

またバージョンアップなど部分的変更に対応するためには、エディタだけでは不十分となる。すなわち DTD に対応したデータベースが必要である。クライアントサーバシステムと組合わせた排他制御などシステム的には複雑になるが、SGML は文書管理ツールと組合わせるメリットは大きい。

4.3 SGML で他国語展開を行なう

SGML はデータベースと組合わせることで、インスタンスのみの変更で他国語展開が可能である。また組版マクロのバリエーションを用意することで、多様な出力形態に対応できる。

組版の自動化も、納期やコストを最優先するクライアントに対して大きなアドバンテイジを持つ。組版マクロで行なえる微調整もかなりあり、フォントやレイアウトの制約をこの段階で補うことも可能である。

HTML: Hyper Text Markup Language
PS: Post Script
PDF: Portable Document Format

5 SGML 導入における問題点

5.1 WYSIWYG でない使いにくさ

作成する側から考えると、SGML はそれほど使い易いシステムではない。デバッグのためのパージング処理が不可欠であるし、レイアウトを確認するためには組版マクロによるレンダリングが必要になる。いずれもリアルタイムな処理はできず、反復操作や待ち時間が生じる。

タグを保持したまま翻訳したり、実際のレイアウトを確認するには SGML エディタを利用してもかなりの熟練を要する。ゼロからの書き起こしでない限り、汎用のエディタは使いにくいといわざるをえない。

5.2 社内規則の厳格な適用の必要

社内的な SGML への部分的に移行にはほとんど意味がない。企画設計から入力のレベルで厳格に規則化を実施しなくては、後工程の作業が複雑になるだけである。本質的に外部の業者に委託できるシステムではない。

そのために SGML 化のためにドキュメントプロセスを一旦停止して、全工程をいっせいに切り替える必要がある。また環境的にも共通の基盤を使わなくてはならず、初期投資がかなり必要である。ドキュメント自体の設計を SGML に適したものに変更できれば理想的である。

5.3 SGML の柔軟性維持の困難性

SGML は標準化されているのは構造のみであり、レンダリング系はまったく考慮されていない。そのため組版ツールによって出力系の機能が大幅に制約を受ける。多くのシステムは組版系の手作業を前提としている。この点も社内システムとしては問題がある。

また DTD を解析する必要があるために、SGML エディタとデータベースとの組合わせは、組版マクロも含めた単一システムになりがちである。これではデータの柔軟性を損なう可能性がある。SGML 本来の意味のオープン化と、現実のこうした囲い込みとの矛盾が問題になる。

WYSIWYG: What You See Is What You Get

6 インターネットとイントラネット

6.1 インターネットの爆発的普及

インターネットは、その淵源を SGML と同じころの 1969 年の米国防総省国防高等研究計画局による ARPAnet にまでさかのぼることができる。しかし AUP と呼ばれる利用規制により、利用者は研究所、大学、一部の大企業に限定されていた。

1990 年からはじまった商業利用の解禁は、世界規模のネットワークの利便性を一挙に一般利用者にまで浸透させることとなった。現在では利用者数は 8,000 万人とも一億人ともいわれている。まさに世界最大の情報ネットワークである。

6.2 インターネットが実現する情報世界

インターネットの情報世界を牽引するのが、WWW とブラウザである。ハイパーテキストを実現できるデータベースである WWW は CERN が 1989 年に開発したものであるが、1993 年に発表されたモザイクと呼ばれる WWW 専用クライアントブラウザによって使い勝手の向上と飛躍的な普及を実現した。

現在はネットスケープが 85%の市場占有率を占めるが、ブラウザはさらに高度な機能

を実現し、その結果日本国内だけでも商用プロバイダは300社近くに達している。マイクロソフト社も1996年度からインターネットにシフトした戦略をとっている。

インターネットの情報は、従来の情報に対する概念を変革した。情報の保有や入手に価値をおくのではなく、どんな情報をどの程度利用しているかがポイントとなる。情報は利用されてはじめて価値が出ることに、オープン化と分散の思想の神髄がある。

6.3 インターネット技術の進化

不特定多数の利用者を前提とするインターネットは、そのボトルネックがネットワークの通信速度にある。そのために、インターネットは低速のネットワークで実現できるアプリケーションが求められる。

それとともにクライアント側の環境の特定が困難なために、それほど過大な負荷を端末側に要求することができない。インターネットは、ネットワークの対応性と通信速度の制約、端末の負担軽減などを維持したまま発展を続けている。

インターネットはその普及度ゆえに、採用されている技術は事実上の業界標準としての地位を獲得している。現在では電子メールやFTP、WWWの他に、動画やサウンドそして電話機能など、マルチメディアの世界まで拡張されている。

ARPA: American Research Project Agency
AUP: Acceptable Use Policy
CERN: Centre Europeen pour la Recherche Nucleaire (the European Laboratory for Particle Physics)
FTP: File Transfar Protocol

7 社内情報システムとイントラネット

7.1 社内情報システムの問題点

従来の社内情報システムは、1970年代に確立したIBM社のSNAプロトコルによるホストシステムが主流を占めていた。メインフレームを核とした集中管理方式は、コストやシングルベンダーへの依存など様々な問題をはらんでいた。

そのために80年代のUNIXワークステーションの台頭により、クライアントサーバシステム(LAN)に移行した。しかしオープンを前提とした分散システムは、逆にユーザ側にリスクを強いることとなり、社内運用コストはむしろ増大する傾向にある。

また分散システムの結果としてクライアント側に過大な負荷を強いることともなり、クライアント数の増大と共に導入や運用管理のコストが増す。システムごとにクライアントGUIが異なるので、設計や教育にも配慮が要求される。

この結果として間接部門である社内情報システム部門の更なる強化と、リスク回避のために特定のベンダーへの依存が増すという状況にあった。すなわち情報部門とベンダーによる情報の独占が、企業単位で進行することになったのである。

7.2 イントラネットコンセプトの登場

インターネット技術を利用した社内情報システムの発想は以前からあった。しかしインターネット技術の発展と、ビジネス利用での普及の頭打ち感が、これに拍車をかけた。1995年末に米国の各雑誌や市場調査機関にイントラネット(INTRANET)の名称が始めて登場して以降、翌年には日本でも盛んに論じられるようになった。

イントラネットを一言で表現すれば、社内ネットワークへのインターネット技術の導入である。ネットワーク技術や利用技術自体は決して新しいものではないが、事実上の標準であるという点で、社内情報システムが抱えていた諸問題を解決できる可能性を持っている。

イントラネットではシステム設計における情報システム部門の負荷を軽減でき、WWWブラウザをクライアント側に利用するために習得も容易である。また特定のベンダーに依存することもなく、ネットワーク技術自体も高度でない。

7.3 イントラネットの可能性と将来性

イントラネットは、社内LANのプロトコルをTCP/IPに変更することからはじまる。

DNS と HTTP で WWW サーバを構築してデータベースを利用することと、電子メール関連機能を実現する POP3、SMTP、NNTP などの各サーバが最低限必要である。

現在のところイントラネットはグループウェアの完全な代替システムとはならない。しかし標準的なツールを利用できるメリットは大きく、テレビ会議システムなどマルチメディア的な利用法にもリスクの少ない投資が可能である。

またインターネットと同じアーキテクチャーであるために、イントラネットはインターネットとシームレスに接続できるのも大きな特徴である。セキュリティに関しても、ファイアウォールや暗号化など社内情報には十分なレベルでの利用技術がある。

SNA: System Network Architecture
GUI: Graphical User Interface
TCP/IP: Transmission Communication Protocol / Internet Protocol
DNS: Domain Name Server
HTTP: Hyper Text Transfar Protocol
POP3: Post Office Protocol 3
SMTP: Simple Mail Transfar Protocol
NNTP: Network News Transfar Protocol

8 イントラネットでの RDBMS の役割

8.1 WWW サーバとクライアントの機能

電子メール機能と並んでイントラネットを特徴付けている WWW は、サーバとクライアントからなる固定的データベースである。WWW サーバ内のデータは、すべて HTML ファイル化されていなければならない。

WWW では、クライアント側のブラウザは、WWW サーバにリクエストを送り、サーバが返してきたファイルを単に表示するだけである。そのために更新の頻繁な社内情報では WWW サーバのメンテナンスが追い付かないことが多い。

そこで社内のデータベースエンジンをアクティブなバックエンドとして利用し、そこからのデータを HTML に変換して WWW サーバに送信する方法が取られている。これが動的（ダイナミック）な変換と呼ばれるシステムである。

8.2 RDBMS とウェブとの連携とは

WWW サーバへデータベースからデータを送る場合、一括で HTML に変換するバッチ的な処理と、アクセスの都度データベースを立ち上げる処理がある。登録や更新までもイントラネットで行なうには、トランザクション処理を実現しなくてはならない。その場合にはアクセスごとの処理が必要になる。

アクセスごとにプロセスを立ち上げる最も簡便な方式は CGI を介在させることである。WWW サーバから CGI でデータベースを立ち上げることで、「連携」処理が可能になる。しかし 2 秒以内を実現するにはやサーバ API の方が有利である。

この場合、一般のクライアントサーバ方式が採用しているサーバとクライアントの 2 層構造とはならない。WWW サーバを経由してデータのやり取りを行なうという、いわば 3 層ミドルウェア方式に近い姿となる。

8.3 ダイナミックデータドリブン・データウェアハウス

データベースへのアクセスと HTML 変換をアクセスごとに行なう方式では、セッションの確立が困難である。トランザクション処理などでは特定のデータにアクセスが集中することが多く、CGI 方式ではプロセスが数百も立ち上がることがある。

近年ではデータベースを定常業務に利用する以外に、分散した社内データベースを経営戦略決定に利用しようとするデータウェアハウス構想が出てきている。この場合はさらに応答性が問題になると考えられる。

サーバ API のような柔軟性の高い 3 層ミドルウェア方式と、十分な応答性を備えた WWW、そして分散されたデータベースを使ったダイナミックデータドリブン・データウェアハウス構想は、イントラネットの最終的な姿ということができるであろう。

8.4 Java 言語とユニバーサルクライアント構想

将来的には現在単に表示だけの機能しか持たないクライアント端末をインテリジェン

ト化するために Java または ActiveX の利用が考えられる。Java は仮想コンピュータ概念によるバイトコードへのコンパイルを行なうために、プラットフォームに依存しない。

Java ないし ActiveX を実装したブラウザは、問合わせのデバッグと、サーバから返されたデータの編集も可能となる。GUI のカスタマイズやエージェント機能の実現もそれほど難しい課題ではない。

インターネットに依拠したイントラネットでは、クライアントの負荷はあくまでも極小である。NC が議論される基盤もそこにある。インテリジェント化されたクライアントは、ユニバーサルクライアントと呼ぶに相応しい形態になるであろう。

RDBMS: Relational DataBase Management System
CGI: Common Gateway Interface
API: Application Programming Interface
NC: Network Computer

9 イントラネットによる SGML の復権

9.1 イントラネットによる CALS の実現

イントラネットを実際に利用するには、社内文書を HTML に変換する作業が必要である。しかし既存の文書を HTML に変換することは、その後のデータ交換や流用という面で大きなハンディを抱える。

もし社内文書の形式をすべて統一することができれば、それはまさに HTML による CALS の実現に他ならない。CALS 技術情報の標準化戦略は、国際標準・国内標準・政府標準・業界標準に段階分けされているが、規格化の内容をあまりにも細分化しすぎたために時間がかかった。インターネットは、現在の技術やインフラでできることをするという正反対のアプローチをとっている。

手法の優劣はつけられないが、現時点では CALS での優位性を確保するためにインターネット／イントラネットで先に標準を獲得しなければならないというパターンになりつつある。こうした状況において、HTML のスーパーセットとしての SGML の存在意義は大きい。出力系に依存しない構造化言語は、ネットワークというメディアの性格からいっても最適である。

9.2 SGML／HTML ダイナミックリンキング

SGML をイントラネットで利用するには、HTML に変換して利用するのが一般的である。しかし Panorama PRO などのプラグインビューアを利用すれば、SGML インスタンスをそのままクライアントに表示できる。

固定的な HTML かダイナミックな変換かを問わず、SGML を HTML へ変換することはハイパーリンクや情報の切りわけ、レイアウトなどの問題が生じる。SGML データのままであれば、入手したデータを直ちにクライアント側のレンダリング系に組み込むことによって再利用が可能である。

9.3 次世代データベース OODB への期待

しかし、SGML を RDBMS でサポートすると、十分な応答性が期待できないという大きな課題がある。検索エンジンにおいても、同様である。これはもはや構造の問題ではなく、データベース自体の特性によると考えられる。

そこでタグが付けられた SGML インスタンスを完全に活用するために、OODB ないし ODBMS の機能が求められている。まだ完全に実用化された OODB はないが、イントラネットのオープン性を維持しつつデータベースとの連携を実現するためにも、RDB では限界がある。

9.4 PDF とドキュメント管理の今後

インターネットを経由してイントラネット同士を結ぶには SGML のようにネイティブな形でのファイル交換が望ましい。しかし

ネットワーク上の文書配信技術は、SGML のような再利用可能形式にこだわらない面もある。

特に画面表示と印刷出力に留意するのであれば、ファイルの再利用には問題があるが、PDF を考慮する必要がある。とくに依然として紙が支配的であるオフィスの現状を考えると、イントラネットでは PDF が有利である。

既存の DTP ファイルを PDF 化するのは SGML 化するよりはるかに容易である。イントラネットのコンセプトが浸透し SGML のメリットが十分認識されるまでは、一時的に HTML の表現を越えた PDF が普及すると考えられる。

CALS: Computer-aided Acquisition and Logistic Support または Commerce At Light Speed とも言われる
OODB: Object Oriented DataBase System
ODBMS: Object DataBase Management System
PDF: Portable Document Format

10 おわりに

インターネットとはいえ、新しい技術が標準化するには時間がかかる。しかしネットワークでシステム化するには、従来の拡張性を残し、特定のベンダーと社内専門家からの開放が必要である。そのためにイントラネットは最も適していると言える。

イントラネットは、いまある技術を使ってスケーラビリティに富んだシステム構築できる最善の社内情報アーキテクチャである。スケーラビリティは拡張性と訳すが、柔軟性ともとれる。真のスケーラビリティは、オープンな環境で分散処理を実現する以外には達成できない。

イントラネットは社内情報化の統合である。SGML も本来的な意味は情報の統合である。当社は将来的にも有用なドキュメントの形態を模索している。イントラネットが SGML の復権に結び付くように、ドキュメント全般におけるシステム構築とサポートすることが当社の役割であると考えている。

SGML は本来的に長期的展望にたって考慮すべきシステムであり、性急にその効果を検証することのできるシステムではない。またそれなりの「決断」を要求されるシステムでもある。当社の方向性を理解いただき、SGML の可能性の正しい把握をしていただければ幸いである。

以　上