出版社，SGML(XML)と出会う

上野真志*

本稿の目的は、一出版社にとってのSGML(XML)のインパクトと、辞典系の出版物や電子出版において行われた技術的な取り組み、そしてそれらの成果を今後どのように活かしていこうとしているのかを紹介することにある。

1. まずは辞典の組版から始まった

(1) 「タグ」との出会い

出版社にとってのSGML(あるいはXML)との出会いを考えると、CTS (Computerized Typesetting System) の導入を画期として挙げることができる。岩波書店においてそれは、1970年代末のことであった。さらに、辞典をターゲットとして導入したため、ある種必然的に、コンピュータ技術を十全に活用するためのさまざまな工夫を凝らすことになる。

幸いにしてパートナーにも恵まれ、汎用のCTSプログラムではなく、辞典ごとの専用システム(と言っても汎用品をカスタマイズしたものであるが)を開発して用いることができた。このときのポイントは、

- ①データベース指向
- ②マクロ定義

の二つに整理できる。現在から考えると非常にprimitiveなものであったのは事実だが、それでも、出版社にとって「CTSなんて」という疑念を払拭するだけでない、大きな副産物をもたらした。それが「タグ」という概念との出会いである。

(2) 組版ファンクションのマクロ化

CTSは、コンピュータ技術を活版に適用する形で出発した。しかし、周知の通り、コンピュ

ータ技術の適用範囲はそこに留まるものではなかった。まず簡単なものとして「代替約物」というものがある。たとえば、(())と入力されたものを、実際の印刷時には〈 〉のように変換して出力する類である。

さらに、もう一つ。CTSには、活版時代から出版社が伝統的に行ってきた各種の指定(たとえば、字体変更や開け詰めの調整など)を実現するために、「ファンクション」と呼ばれる〈機能文字〉が実装されている。このファンクションの考え方はCTS組版の根幹をなすと言っても良く、先に述べた専用システムは、まずこのファンクションを読み替えることから始まった。

たとえば、ある辞典のシステムでは、ファンクションXが入力されているとき、ファンクションA+B+Cの機能を果たすように定義されている。このようにして実現された「マクロ定義」は、そもそも、パンチャー(入力者)のストローク数を減らし、入力ミスを避けるために考えられたものであるが、定義されたマクロをもとに処理を行う際には、一種の「タグ」として働くことになる。

(3) データベースとしての活用

もう一つ、CTSで辞典の組版を行う際に考えられたのが、データベースへの指向である。

たとえば、ファンクション CS と CE で括ら

* A Publisher meets SGML(XML)

Masashi Ueno, (株)岩波書店 電子出版部, Iwanami Shoten, Publishers
e-mail: ueno@Iwanami.co.jp

れた部分を実際の印刷では印字しない、という約束事をつくり、CS ～ CE の間に、項目(レコード)を管理するための情報を置く、という対応がなされたシステムがある。この囲まれた部分を「項目管理部」と呼び、辞典ごとに、編集方針に対応した形で、さまざまなコード付けを行っている。図1は『広辞苑』での項目管理部の情報の一例である。

- 品詞
- 和語／漢語／外来語
- 項目のジャンル(法律、物理など)
- 語種：季語、慣用句(成句)、故事成語など
- 項目履歴
- ソートキー

図1 『広辞苑』の項目管理部の情報

(4) スタイルの分離

これらの二つをさらに発展させたシステムもある。項目のデータを、項目管理部と項目本文の二つに分けるのではなく、後者をさらに細分化するのである。

単純に言って、辞典の各項目は、見出しと解説文に二分できる。さらに、見出しの中での情報(かな／漢字表記など)の組体裁や解説文の情報(品詞や活用形、用例や生没年など)の出現順なども厳密に決められている。つまり、項目本文を適切に細分化するという事は、適切な意味のまとまり(論理的なブロック)に分けるということになる。

さらに、辞典では、これらの意味のまとまりは、ある特定の組体裁を伴っている。つまり、適切にブロック化が行われれば、自ずからブロック内の組版方法は一意に定まる。スタイルと本文データの分離がなされたシステムの誕生である。

2. 電子出版との関わりで

このように見えてくると、SGML(XML)の利点として考えられるもののうちいくつかは、辞典

ごとの専用システムでも実現されていたように思える。少なくとも冊子体(書籍)で出版する限りにおいては、ある水準を満たしていた。ここから、SGML(XML)に目を向けさせる方向に働いたのが、電子出版である。

『広辞苑』によれば、電子出版とは「編集・印刷の過程をコンピューターによって管理した出版のこと。また、CD-ROMなどのニュー・メディアで大量の情報伝達を行うこと」とある。当面話題となるのは、もちろん後半部分の定義であるが、前半にも関係する。

(1) 出発点となる「ソース」は何か

電子出版が立ち上がり始めたころ、「One Source, Multi Use」ということが盛んに唱えられた。その背景には、やはりCTSの普及がある。ただし、振り返ってみるとき、その実態は図2のようなものであった。

図2 初期の One Source, Multi Use

この形態で作業を進行するときの問題点は明らかである。すなわち、

- 冊子体用データが校了(データ・アップ)にならない限り、電子出版用のデータが作成できない。
- 個々の電子出版物ごとにマスターが存在し、一元管理が困難である。
- 改訂が行われた際には、電子出版用データ

が一から作成し直しになる。

という点である。これらは、すべてコストの問題として跳ね返ってくる。

この問題にわれわれがぶつかったのが、『広辞苑第四版』の CD-ROM を作成したときである(『広辞苑』は 1987 年に第三版を CD-ROM 化していた)。第四版の冊子体の刊行は 1991 年 11 月であったが、電子出版物(電子ブック版/CD-ROM 版)の刊行はそれから 1 年から 1 年半後のこととなった。

(2) SGML を使う

問題を解決するためにわれわれが選んだ方法が、「冊子体と電子出版に共通のマスターを作成する」ということであり(図 3)、その際に用いた技術が SGML であった。すなわち、

- ①データはすべて SGML 形式で記述し、基本的な校正は、その SGML データに対して行う。
- ②冊子体の組版や電子出版物の作成は、それぞれ専用プログラムを通して変換することを基本とする。すなわち、最終段階まで、データを一元管理する。

ことを基本とし、1993 年に大日本印刷と共同で研究を開始した。

図3 目標とした One Source, Multi Use

SGML 化の出発点は、各種辞典の項目記述構造の洗い出しである。先に述べたように、辞典においては論理ブロック化が比較的容易であり、このブロック化がなされれば、組版方法も決定できる。しかし、一方で、これまでの辞典(冊子体)は、人間が読めば判断できることを前提としていた(=機械では一律に判断できないことがある)り、限られたスペースの中で情報を的確に伝えるために例外処理を行ったり、ということがあった。これらを考慮しつつ、着地点を見いだす作業はかなりの困難であった。

もう一つ問題となったのが、校正(データ修正の指示)方法である。辞典に限らず、出版物の校正は、ゲラ(という印刷物)で行われる。このときの修正指示は、主として、データの内容の誤りについて行われる。一方、SGML 化した場合は、SGML タグの付与方法の誤り(フォーマットエラー)の有無をチェックする必要もあるが、これらは通常、モニタ(SGML ソースのプリントアウト)で行われる。そこで今回の作業では、「モニラ」(モニタ+ゲラからの造語)なるものを案出し、データ内容の誤りと SGML のフォーマットエラーの双方を、比較的無理なく行えるようにした。

これらの研究成果をもとに、『広辞苑第五版』の作業を開始し、1998 年 11 月に冊子体と CD-ROM 版の同時刊行を行った。また、ここで作成された SGML データは、サードパーティへの『広辞苑』のライセンス時にも用いられている。

(3) 次ステップへの課題

今回の一連の取り組みで明らかになった課題としては、

- ①SGML タグの(無原則な)増殖
- ②リテラシー
- ③記述に用いた言語

という 3 点が挙げられる。

まず①について、今回の『広辞苑』の作業では、時間的制約から、すべての必要事項を洗い

出した上で、さらに繰り返してテストを行う、ということができず、作業を進行させつつ決定していく方法をとった。そのため、組版プログラムの制約からタグを設けなければならなくなったり、属性値の変更／追加で対応できるものが別のタグになったり、といった事態がおこっている。これらは、今回の作業上はそれほど問題にはならなかったが、全体としてフォーマットの見通しを悪くするために、将来的には問題となる可能性がある。

②については、辞典の改訂がチームを組んで行う作業であることに起因する。たとえば DTD 一つとってみても、それをチームの全員が解読できるとは限らない(少なくとも出版社においては、そんなことを期待するほうが間違っている)。今回は DTD を“パラフレーズ”したドキュメントを用意したが、それでもタグの意味の理解にばらつきを生じ、結果として人間／機械の双方で作業上の負荷となった。対応には、チーム内での共通理解を進めることが必要であるが、なかなか容易ではない。

さらに、もう少し根幹の問題として考えられるのが③である。すなわち『広辞苑』の記述は、XML ベースではなく SGML ベースで行われている。今後の言語仕様の発展や、周辺環境(パーサなどのプログラム類)の整備状況を考えると、いずれ XML に対応した形で記述し直すことも必要になるだろう。

3. 辞典だけが出版物ではないのだが

さて、このように研究や実践が進んでくると、「次は辞典以外の書籍をターゲットにして」という考え(あるいは「要請」)が出てくる。これは一見もっともな考え方であるが、実は非常に乱暴な括り方でもある。それはたとえば、学術書と文芸書と美術書(写真集)の違いを考えてみれば、明らかであろう。

十把一絡げに「書籍」というのではなく、もう少し絞り込んだターゲットを設けて、フォー

マットの研究を進めるのが適当である。昨年、日本電子出版協会によって交換用フォーマット「JepaX」が提案されたが、そこでは新書・文庫・一般書など、“レイアウトが単純で文字主体の出版物”をターゲットとしている。岩波書店においては、岩波新書がターゲットとして妥当である、という判断が成り立つ。

以下は、岩波新書の XML 化を目標とした場合の課題などを列挙したものである。

(1) データの利用方法は？

XML 化したときのデータ利用方法として有力なのが、コンテンツのダウンロード販売(オンライン出版)である。著作権保護の仕組みなどはとりあえず置くが、各種の電子書籍フォーマットは概ね XML を指向しているのは事実。また、スタイルシートと XSL プロセッサの使い方によっては、最近流行のオンデマンド出版にも益するところ大であろう。

さらに、岩波新書をターゲットとしたときには、辞典のように単体で使う以外の用途が想定可能である。そもそも、岩波新書には、超大項目主義の百科事典的な観点から企画／刊行されたものが多い。したがって、複数の岩波新書をまとめた形でデータベース化し「知の体系」を提供する、というのも有力な方法である。

付言すれば、自然言語処理の研究用のソースとしても XML 化したデータ類は有用なはずである。

(2) 交換用フォーマットの条件

こういったデータの用途を考えたとき、岩波新書をターゲットとした中間フォーマット(交換用フォーマット)の必要性は明らかである。ただし、そのフォーマットは、交換可能なフォーマットの多さを保証するだけでなく、「データを、出来る限り低コストで、継続的に作成できる」という条件をも満たすものでなければならない。

前述の利用方法で考えてみよう。ある時期までに岩波新書 50 冊を XML 化するという事は、コストとマンパワーの問題に帰着させることで、比較的实现が容易である(それでもかなりの規模になるが)。それによって、オンライン出版はとりあえず可能となるし、「知の体系」の一部を提供することにもなる。しかし、毎月刊行される岩波新書のラインナップを取り込みつつ、さらなる発展形を望むことにはつながらない。

周知の通り、昨年提案された JapaX は、交換フォーマットとしての有効性(他フォーマットへの変換など)が、ある程度すでに保証されている。現在は実証実験フェーズにあると考えられるが、先の条件の検証には残念ながらもう少し時間がかかりそうである。

出版社は伝統的に生産プロセスのある部分を全面的に他社に依存している会社である。今回の事例で言えば、CTS のデータを実際にメンテナンスしているのは、印刷所である。たとえば岩波新書の印刷には、複数の印刷所が関わっている。さらに、そこで用いられているのは汎用の CTS プログラムである。

先の条件を満たすということは、すなわち、これらの印刷所において、現在の環境を大きく変更することなく、比較的に作業が行えるようにすることと言い換えられる。

(3) 汎用 CTS の問題点

現在われわれは、先の条件を満たすようにするための予備調査を開始したところである。すでに述べたように、岩波新書の組版に用いられているのは、汎用の CTS プログラムである。まず、今回の予備調査では、XML(JapaX)化の出发点として想定できる、「MS-DOS テキストの出力」という CTS プログラムの補助機能を対象にした。

これまでに明らかになった問題点の一つに、「CTS データ⇔DOS テキストの完全変換は保証されていない」ということがある。ファンクシ

ョンのうち特定のものは、JIS 内文字の組合せ(たとえば「@改行」や「■1F5B」など)に置換される形で変換されてくるが、すべてのファンクションが対象になってはいない。逆も同様である。

これはたとえば、XML ファイル+XSL プロセッサで作成したデータが組版システムに取り込めない、すなわち、交換の相手先として組版システムを想定できないことを意味する。この原因は、明らかに

- CTS プログラムが、データの二次利用をふまえた設計になっていないこと。
- CTS メーカーに、データの構造を公開する意志がないこと。

という 2 点に集約される。したがって、利用者であるわれわれには手が出せない部分であり、メーカーの方針変更/バージョンアップを待つしかない。CTS プログラムに XML 形式でのファイルの出力機能が備わる可能性はどれほどあるのだろうか。

(4) その他の課題

CTS プログラム側に XML 形式でのファイル出力機能が無い以上、既刊書の XML(JapaX)マークアップにあたっては、組体裁(ファンクション類)を手がかりにするほかない。そして、ある組体裁がどのような XML(JapaX)のタグ(=論理的要素)にあたるか、ということは基本的に出版社側が指定すべきことである。

したがって、「正しくマークアップされた」XML(JapaX)データを作成するためには、「どのようなマークアップを施した XML(JapaX)データに変換して欲しいのか」を示す「変換指示書」を、変換対象となる書籍ごとに作成する必要が生ずる。ただし、これは事実上不可能であり、前述の条件からも外れることとなる。

現実的な解としては、「指定しなくても正しくマークアップできるものを増やしていく」ということになる。現在、具体的に、このような

暗黙の前提を増やすべく、洗い出し作業を行っ

ているところである。