連載

XML 誕生秘話

村田　真

日本 IBM（株）東京基礎研究所
国際大学研究所
mmurata@trl.ibm.com

3　国際化

□■まえがき

WWW 全体の国際化のためには，どんな自然言語で書かれた情報も XML で扱えることがきわめて重要である．XML が一部の言語しか扱えなければ，XML に基づくフォーマット（たとえば HTML の後継である XHTML や Web サイトについてのメタ情報である RSS[1]）もやはり一部の言語しか扱えない．本稿では，XML がどのように国際化されたかを振り返ってみたい．なお，XML の国際化と日本語プロファイルの詳細については，文献 2）も参照されたい．

□■ SGML の国際化と HTML の国際化

XML の制定以前に，SGML の国際化と HTML の国際化があった．これらを抜きにして，XML の国際化を語ることはできない．

XML の前身である SGML[3] は，ISO 2022[4] を基盤として多くの文字コードに対応する方針をとっていた．しかし，これは 2 つの理由により機能しなかった．第 1 に，SGML 規格は ISO 2022 の不正確な理解に基づいていた．第 2 に，SGML を処理するソフトウェアは，ISO 2022 非準拠のシフト JIS などを扱わざるを得ない．その結果，国際化に関する限り，現実のソフトウェアは SGML 規格を無視することになった．文字コードに関連する機構である SGML 宣言はほとんど実装されず，シフト JIS を含む各種の符号化方式が場当たり的に実装された．SGML 規格委員会の無能を言い立てるのは易しいが，ISO 2022 の側にも，仕様が複雑である，不適合な実装が横行している，扱えない文字コードが多いなどの問題点がある．

文字コード以外にも，SGML の国際化には問題点が多い．たとえば，タグ名や属性名として使用できるのは US-ASCII の文字に限られるという問題点，番号によって文字を参照する機構（XML にも引き継がれた文字参照）が 16 進数を扱えないという問題点などがあった．これらの問題を解決するため，Rick Jelliffe が SGML の拡張を検討した[5]．そこでの成果として，タグ名・属性名についての拡張[6] は SGML の改定に取り入れられた．

HTML も最初から国際化されていたわけではない．一応 SGML に準拠していたが，HTML 2.0[7] は元々 ISO 8859-1 しか扱えないものであった．HTML の国際化については，Gavin Nicol の提案[8] が出発点となった．HTML の document character set を Unicode[9] に固定するという方針，シフト JIS や ISO-2022-JP は Unicode の一部の文字だけを表現可能な符号化方式として許容するという方針は，彼によって打ち出された．この方針は，HTML の国際化のための RFC 2070[10] として発行され，後に HTML 4.01[11] や XML 1.0 にも受け継がれた．

□■ XML の国際化

普通，国際化を専門としないグループは，国際化を苦手として敬遠するものである．日本国内でさえ，すべての分野の専門家が日本国内の文字コードに通暁している訳ではないことを考慮すれば，国際化が敬遠されるのは止むを得ないことかもしれない．

しかし，XML を制定した W3C XML WG は，きわめて国際化に熱心なグループであり，膨大な工数を国際化に費やした．メンバのうち，James Clark および Tim Bray は日本語を含む非欧米言語の文字の実装経験があった．それ以外のメンバも，XML の国際化は WWW 全

体の国際化に大変重要であることを認識していた．前述した Gavin Nicol と Rick Jelliffe も W3C XML SIG に当初から加わっていた．

XML 1.0 の国際化は，SGML および HTML のそれを引き継いでいる．すなわち，Unicode を基準としつつ，シフト JIS や ISO-2022-JP などは Unicode の一部の文字だけを表現可能な符号化方式として許容されている．符号化方式の判定方法は，XML 1.0 および RFC 3023[12] に明確に規定されている．タグ名としては US-ASCII 以外の文字が利用可能であり，文字参照では 16 進数が利用可能である．

この結果，XML はその誕生直後から，日本語を含む多くの言語の文字を扱うことができた．XML の国際化に問題がないわけではないが，他のほとんどの技術が US-ASCII や ISO-8859-1 だけに限定して出発するのと比べれば雲泥の差がある．

□■ Unicode について

XML の国際化において，HTML と同じ方針を採用することは 1996 年 9 月の段階ですでに決定していた．つまり，XML の document character set は Unicode であり，シフト JIS や日本語 EUC などは Unicode の一部の文字だけを表現可能な符号化方式として許容されていた．

私が XML のドラフトを最初に見て，Unicode が基準であると知ったのは 1996 年 11 月である．このとき私は強い嫌悪を感じた．Unicode といえば国内では蛇蝎のごとく嫌われていた頃である．

しかし，XML は Unicode 以外にも，その他の符号化方式をオプションとして認めていた．つまり，シフト JIS の XML 文書や日本語 EUC の XML 文書も認められているのである．Unicode が基準であることの害を危ぶみつつ，とりあえずは目をつぶることにした．

まず，最初に取り組んだのは，シフト JIS, 日本語 EUC, ISO-2022-JP についての明確化である．また，当時進行中であった JIS X 0213[13] に対応可能なことも保障しておきたいと考えた．これらの点について，日本国内の SGML 関係者と議論して多くのコメントを W3C XML SIG に送付した．私が W3C XML WG に加わる以前のことである．このコメントには議長の Jon Bosak から感謝のメールがあった．きわめて早い時期に日本からまとまった意見があったのは，XML を推進する側にとっても嬉しいことだったのだろう．

しかし，送ったコメントの多くは XML 1.0 に取り入れられなかった．日本語についてだけ詳細な情報を XML 本体に盛り込むわけにはいかないというのが理由であった．代わりに，それらの情報をまとめた仕様を別に作ることを打診された．これが日本語プロファイル（JIS TR[14] および W3C の Technical Note[15]）を作ったきっかけである☆1．なお，日本人は日本語のことだけ考えて不満を並べ，国際化を推進する側は，より多くの言語に対してたとえ貧弱であっても統一的な解を与えようとするという傾向があるように思う．このときのコメント審議はその典型である．

XML が Unicode を基準として用いていることには，その後も引っかかっていた．Unicode を基準とすることに反対するメールを書きかけたこともある．しかし，そのメールを出すことは結局なかった．むしろ XML から Unicode をきちんと扱えるようにすることに心を砕いた．たとえば，W3C XML WG において私が最初にしたのは，UTF-16 への賛成票を投じることであった．

私が「変節」したのはなぜだろうか．それは，より多くの言語を扱える現実的な解がほかに存在しないからである．W3C XML WG の一員となった私には，世界の XML ユーザに多少の責任がある．どの言語に対しても理想的には同等の機能，少なくとも最低限の機能を提供するよう尽くさなければならない．ISO 2022 はすでに SGML の国際化において失敗しており，XML コミュニティではまったく見限られている．統一的な解がほかにない以上，Unicode 以外に選択肢はない．Unicode についての情報は，Unicode 規格書を中心として整理されており，仕様制定にも開発にも便利である．これに比肩できるような代案はまったく存在しない．

前述したように，XML は誕生直後から日本語文字を扱うことができた．これは，Unicode を採用したことの大きな利点である．Unicode が大きな問題を抱えていることは私も同意するが，Unicode の功績も大きいことは事実である．

□■ WWW の国際化における日本人

私はほぼ 2 年半に渡って W3C での XML 制定活動にかかわってきた．その過程で，W3C のメンバ，XML 関係者，Unicode 関係者などとメールや口頭で議論をし，Unicode/HTTP/MIME/SGML/HTML などの仕様に目を通してきた．国際化のために営々として積み重ねられてきた努力に触れることができたのはきわめて貴重な経験であった．

XML の仕様制定活動にかかわる以前は，WWW の国際化において中心的な活動をする日本人はいくらでもいるものと思っていた．当時，Unicode 批判を繰り広げる

☆1 日本語プロファイルは何ら問題を解決したわけではないが，どのような問題があるかを国際的に知らしめた点に多少の存在意義がある．

論客は国内には少なくなかった．国際の場でも積極的な発言をする人は多いのだろうと想像していた．ところがXMLの仕様制定活動に深く携わるようになっても，国内の論客の名前を見ることは稀であった．私の知る限り，Unicodeに反対する理由も，国際の場にはほとんど伝わっておらず，感情的な攘夷論であるとしか受け止められていない☆2．では，日本人の姿が見当たらないのかというとそうではない．国内の論客とは別に，国際的な仕様制定活動で活躍する日本人がいるのである．彼らが国内で寡黙だというわけではないが，国内より国際に重点が置かれていることは間違いない．逆に，国内で有名な論客の多くは，国際的な仕様制定の場には登場しない．国際派と国内派がほぼ完全に分離しているのは残念なことである．

国際の場では数多くの人がWWWとインターネットの国際化のための努力を重ねてきている．筆者は1997年以降の活動にしか関与していないが，それ以前にも膨大な努力があったことは資料から窺える．最初はISO-8859-1にある文字しか扱えなかったWWWが，多くの言語をまがりなりにも扱えるようになりつつあるのは，多くの人の献身的な努力の結果である．そこには，富める国の主要言語だけを扱うのではなく，すべての言語に対して理想的には同等の機能，少なくとも最低限の機能を提供しようという精神が存在する．

国際化には多くの問題が伴うことを承知の上で言えば，日本国内の事情だけを考えた議論を日本国内だけで繰り返しても得られるものは少ないように思う．他の言語を考慮した議論を国際の場で展開するのでなければ，インターネットの仕様に実質的な影響を与えることも，他国の尊敬を得ることもないと私は考える．

□■問題点

本稿では触れなかったが，Unicodeにはさまざまの問題点がある．XMLの制定以降（すなわちUnicode 2.0以降）に付け加わった問題点も多い．数式用文字や言語タグをはじめとして文字とは認めがたいものが数多く追加された．ただし，これらとXMLの併用は禁止されている[16]．各国が漢字を追加・変更したことによってCJK Unificationの問題点も顕在化した[17]．

XML制定以降の国内における重要な動きとして，JIS X 0213の制定，表外漢字字体表，住民基本台帳のための統一文字コード，人名漢字の追加，JIS X 0213の改定などがある[18]．当事者ではない私にも，文字コード規格を保守・拡張していくことには，万里の長城にも比すべき困難さがあることが伺える．数多くの漢字に対して，広範囲の人が合意する包摂規準を長期間に渡って維持することは不可能ではないかと思う．

今後は，符号化文字集合への拡張を繰り返すのではなく，XMLなどの上位レイヤにおいて漢字を扱うことを検討すべきだと考える．文字と番号とを同一視するという方針は，コンピュータが非力であった頃には止むを得なかったが，今となっては有害な呪縛かもしれない．符号化文字集合ではない解を模索しているCHISE project[19]に期待している．

参考文献

1）RDF Site Summary（RSS）1.0, http://web.resource.org/rss/1.0/（2000）.
2）村田　真：XMLの国際化と日本語プロファイル，コンピュータソフトウェア，Vol.16, No.4（1999）.
3）Standard Generalized Markup Language（SGML），ISO 8879（1986）.
4）Character Code Structure and Extension Techniques, ISO/IEC 2022,（1994）.
5）Jelliffe, R.: Extended Reference Concrete Syntax for SGML, http://www.sil.org/sgml/ercs.txt（1995）.
6）ISO 8879 TC for Extended Naming Rules for SGML（1996）.
7）Berners-Lee, T. and Connolly, D.：Hypertext Markup Language-2.0, IETF RFC 1866（1995）.
8）Nicol, G.: The Multilingual World Wide Web, http://www.oasis-open.org/cover/nicol-multwww.html（1994）.
9）The Unicode Standard（Version 3.0），Addison-Wesley（2000）（ただし最新版は4.0）.
10）Yergeau, F. et al.: Internationalization of the Hypertext Markup Language, IETF RFC 2070（Jan.1997）.
11）Raggett, D. et al.: HTML 4.01 Specification, W3C Recommendation（1999）.
12）Murata, M. et al.: XML Media Types, IETF RFC 3023（2001）.
13）日本工業規格，情報交換用符号化拡張漢字集合JIS X 0213:2000, 日本規格協会（2000）.
14）標準情報（TR），XML日本語プロファイルTR X 0015:1999, 日本規格協会（1999）.
15）Murata, M.（ed）：XML Japanese Profile, http://www.w3.org/TR/japanese-xml/, W3C Technical Report（2000）.
16）Dürst, M. and Freytag, A. :Unicode in XML and other Markup Languages Unicode Technical Report #20 W3C Note（2002）.
17）川幡太一：新ISO/IEC 10646とUnicodeの漢字を検証する，漢字文献情報処理，第2号，漢字文献情報処理研究会（2001）.
18）小形克宏：文字の海，ビットの舟，http://internet.watch.impress.co.jp/www/column/ogata/（2000-2003）.
19）CHISE project, http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise/index.html.ja.iso-2022-jp（2000-2003）.

（平成15年5月8日受付）

☆2 XML1.0が制定された頃，「お前がXML WGに入ったとき，たいへん心配した」と言われたことがある．