XML の骨抜き利用法

アジア・アフリカ言語文化研究所データベースの例

豊島正之 (東京外国語大学アジア・アフリカ言語文化研究所) [1]

2001 年 10 月 27 日 九州大学文学部

XML pro/con – XML で書く文献学的データ

古典学の再構築 – 情報処理 (A03) 班主宰研究集会

1 tag は有用・構造記述は無用

1.1 SGML は 、XML になって良くなったか

XML は、SGML を単純化したものだから、SGML の欠陥の多くを継承している。

XML が SGML を改善したのは、処理プログラムの負荷の軽減が主であって、文書作成・交換に関わる問題は、本質的に SGML の持つ問題を解決していない。

(1) 文書は、階層的な要素によって構成されるとする前提。

要素の境界によって文書の境界を兼用する「タイル貼り方式」の文書モデルで、異解釈、欄外注記 (marginaria)、誤写･虫損･欠損 (lacunae)、掛詞、渡リゼリフなど、解釈によって異なる構成を共存させる事が、困難 (SGML[CONCUR])、又は不可能 (XML)。

又、empty element は、全てこの構造の例外となるが、実際の文献翻刻にはページ付けは必須の要素であるし、他に、挿画等の非構成的要素も頻繁に現われる。

そもそも、ページ付け等は、本来は「ページ」「丁」(Octavo 等の)「折」といったブロックを成す element のマークである。しかし、これらの外形のブロックは、「section」「paragraph」といった内容のブロック element と (殆どの場合) 階層構造をなさない別系列のブロックとなり、XML の前提とする「唯一本の木構造」では表現出来ない。このため、外形ブロックの記述の替りに「ブロックの境界」という empty element を立てているだけであり、この様な empty element の利用法は、欺瞞である。

又、クロストークの様な非線条的 (serial) なデータ構造の表現には、empty element で文書中にアンカーを打ち、そこに ID attribute を振って相互参照するという`<a name="ID">` `<href a="ID">`型の表現が用いられるが、これは、empty element の abuse である。

(2) 意味は形式 (構造) が構造内の位置 (価値 valeur) によって付与するものであり、ラベルや内容の持つ意味には関わらないとする、構造主義的な形式・意味分離モデルの本質的な問題。

(3) attribute に構造が存在しない。

(4) element タグの意味交換が不可能。(後述)

[1] mtoyo@aa.tufs.ac.jp

(5) 文書の入れ子が出来ない。

上位文書に読み込まれるためには、独立文書として適格・well-formed で**あってはならない**。

(6) 名前に局所性・スコープが無い。

といった問題[2] は、そのまま XML に継承されている。

SGML の欠陥のうち、名前の局所性は、XML は「名前空間」を持ち込んで解決しようとした。これは、文書の element/attribute にユニークな前綴りを付ける事で全てを固有名詞化して衝突を避けようとしたものである。通常のプログラム言語での「名前空間」は、当該ブロック内での identity を保証して情報共有を許すものの筈であるが、XML の「名前空間」は、衝突を避ける余り異化 differentiation を最優先事項とし、identify の機能は捨て去って仕舞った。(後述する様に、SGML の attribute は 複数 element 間での情報共有が出来ないが、名前空間の適切な設計の下では、これを可能に出来たかも知れないのである)。この結果、XML 文書の validation は時に直観に反する結果を与え、実質的にほぼ不可能になった。Bourret(2001) は、namespace を使って valid 且つ conforming な XML 文書を書く為には

- Use one prefix per XML namespace.
- Do not use the same prefix for more than one XML namespace.
- Use at most one default XML namespace.

The latter three guidelines guarantee that prefixes are unique.

せよとするが、折角 namespace 仕様を設定して置きながらこうした「自主規制」が必要になるとは、仕様の失敗としか思われない。namespace 仕様が最優先事項とした identifier のユニーク性すら、実は**保証出来ない**様である。

1.2 アジア・アフリカ言語文化研究所のデータベースでの骨抜き XML

アジア・アフリカ言語文化研究所 (以下「AA 研」) で公開中のデータベースの内、豊島が管理している下記は、ソースは全て XML で記述してある。これは、tag の well-formedness の検証、要素の抽出 (DOM) 等に既存の XML ソフトウェアが便利である、というだけの理由であり、文書構造の検証等は一切意図していない。

そもそも、文書構造による構造的意味の付与自体、こうした文献学的な翻刻では殆ど無意味である。DTD 等による構造的意味の付与とは、既に理解された構造に基づく意味をパタン化し、複数の文献をそのパタンに基づいて範列化して把握しようとするものである。一方、文献学的研究とは、(正面切って言うのも恥かしいが) その文献の構造自体が研究対象の一つであって、データ化に「構造が既に理解されている」という前提を措く事自体が誤っている。

以上から、文書構造による構造的意味の付与は行わない。従って、(文書構造の検証の役にしか立たない)DTD も作成しない。文献学的な翻刻で DTD が必要になるのは、element の相互関係が余程複雑になって検証困難になるケースか、attribute が極めて多数ある場合等に限られるのではなかろうか。AA 研のデータベースの設計は単純を旨としているので、このどちらにも該当しない。

簡単に言えば、AA 研の下記のデータベースでは、XML の「構造的文書」特性を全て骨抜きにし、ソフトウェアのスーツが準備された便利なタグ記述言語、という側面のみを利用している。

1.2.1 tag が欲しい実例

(1) 三省堂「言語学大辞典」データベース `http://irc.aa.tufs.ac.jp/SLEDB/`

版元三省堂の全面的な協力を仰ぎ、科学研究費補助金を得て、AA 研情報資源利用研究センターが進めている全文検索マルチメディア・データベース。現在、主として権利関係から、ごく一部を公開中。

データベースの素材の性質上、音素、形態素等、語形が短いものを検索したい事が多いが、単純な検索では、これは通常語の「ゴミ用例」の中に埋没して仕舞う。そこで、元データのコーディングに際して「eadem lingua (当該言語)」タグを設定し、記述対象となっている言語のみが選定出来る様にした。

[2] これらに就ては、出雲・豊島 (1993)、豊島 (1992)、豊島 (1994) に詳述した。

eadem lingua (当該言語) タグで「q」を検索する指定。

SLE DB

mtoyo@aa.tufs.ac.jp, 2000

SLE 「三省堂言語学大辞典」本文検索

last update, 2000/09/11

本データベースは、 株式会社三省堂の許諾を得て、 東京外国語大学アジア・アフリカ言語文化研究所 情報資源利用研究センターが作成・管理しているものです。 各項目・記事の著作権は、それぞれの項目・記事の著者にあり、 データベースの著作権は 東京外国語大学アジア・アフリカ言語文化研究所 情報資源利用研究センターにあります。

このデータベース又はその検索結果の全部又は一部は、 学術目的での私的使用に限って、複製が許諾されています。

general keywords	一般的な文字列(正規表現可)	
eadem lingua	当該言語の用例(正規表現可)	q

空白で区切ると AND、|(縦棒)で OR、!(感嘆符)で否定。
例　クメール　語(派|族)　　「クメール」と「語派」又は「語族」を含む記事
例　接(頭|尾)*[語辞詞]　　「接頭語」「接尾辞」「接辞」「接頭詞」…を検索

100 件まで結果を表示(0 で無制限)

検索開始　やり直す

mtoyo@aa.tufs.ac.jp

http://irc.aa.tufs.ac.jp/SLEDB/ [01/10/23 9:18:45]

単純に「q」を検索すると、ゴミが多く、音素、形態素等の用例は埋没する。

SLE search result for (text(q))

2001/10/23:09:22:21 generated by `SLE-s.cgi', do not edit

SLE search result for (text(q))

62 例中 62 例表示

1. TEXT/1-1035L1 オーストロアジア語族 英 Austroasiatic (坂本恭章) q

   Logan, James Richardson (1856), “ Ethnologyof the Indo-Pacific islands", Journal of In-dian Archipelago, New Series 1 (Singapore)Forbes, C. J. F. Smith (1878), “ On the con-nexion of the M&o09;ns of Pegu with the Kolesof Central India", Journal of the RoyalAsiatic Society of Great Britain and Ireland,New Series 10 (London)Keane, Augustus Henry (1882-83), “ Des rap-ports ethnologi ques et linguistiques des racesindochinoises et indo-pacifi ques", Annales del'Extr&e03;me-Orient 5 (Paris)Terrien de Lacouperie, Albert Etienne JeanBaptiste (1885-87), The languages of Chinabefore the Chinese (Trans.

2. TEXT/1-1035L1 オーストロアジア語族 英 Austroasiatic (坂本恭章) q

   N. F. 5(Braunschweig)Hevesy, Wilhelm von (1930-32), “ On W.Schmidt's Munda-Mon-Khmer comparison(Does an 'Austric' family of languagesexist?)", Bulletin of School of OrientalStudies 6 (London)Haudricourt, Andr&e01;-G. (1954), “ De l'originedes tons en vi&e03;tnamien", Journal AsiatiqueCCXLII (Paris)Pinnow, Heinz-J&u05;rgen (1959), Versuch einerhistorischen Lautlehre der Kharia-Sprache(Otto Harrassowiz, Wiesbaden)----音韻研究の集大成ともいうべき大作.

3. TEXT/1-1117R1 オン語 Ong (坂本恭章) q

   [参考文献]Ferlus, Michel (1974), “ La langue ong, muta-tions consonanti ques et transphonologisa-tions", Asie du Sud-Est et Monde Insu-lindien V, 1 (CeDRASEMI, Paris)[参照] カトゥ語派,オーストロアジア語族(坂本恭章)

4. TEXT/1-1133L1 カオ語 Cao (坂本恭章) q

   [参考文献]Ferlus, Michel (1974), “ D&e01;limitation des groupeslinguistiques austroasiatiques dans le centreindochinois", Asie du Sud-Est et Monde In-sulindien V, 1 (CeDRASEMI, Paris)[参照] オーストロアジア語族,カトゥ語派(坂本恭章)

5. TEXT/1-1133L2 カオ諸語 Khao (坂本恭章)

   北ヴェトナムのソンラ(Son-La),ライチャウ(Lai-Chau),北ラオスのポンサリ(Phong Sali),上ナ〓タ(Haut-Nam Tha)などに分布している,サー・スアク(Xa Xuak),カー・クゥン・リム(KhaQuang Lim),カー・ビット(Kha Bit),サー・カオ(X&a01; Khao)などの言語からなる語群で,オーストロアジア語族のモン・クメール語族,クム語派に属する.

http://irc.aa.tufs.ac.jp/SLEDB/SLE-s.cgi (1/10) [01/10/23 9:22:24]

「q」を音素目録に持つ言語は、(公開中のデータベースには) 一つしか無かった。

SLE search result for (SL(q))

2001/10/23:10:55:19 generated by `SLE-s.cgi', do not edit

SLE search result for (SL(q))

1 例中 1 例表示

1. TEXT/1-1430L1 クア語^1 Cua (坂本恭章) q

   語末の子音に,後に声門閉鎖音を伴う /-w q, -y q/, 後に気音を伴う /-wh, -yh, -lh/がある.

SLE project : sle@aa.tufs.ac.jp

(2) 「国語学」全文データベース http://jcs.aa.tufs.ac.jp/SJLDB/

国語学会データベース委員会が、科学研究費補助金を得て進めている機関誌「国語学」の全文検索データベース構築。AA 研情報資源利用研究センターが、技術的側面から支援している。現在、主として権利関係から、「展望」記事の一部を公開中。

国語学では、資料名が、そのまま研究術語名や (当該分野の) 専門雑誌名になる事が多いため、単純な検索では、これらが混雑して、「ゴミ用例」が増える。又、「展望」(サーベイ) 記事では、展望記事の著者名・論文名等と、展望 (論評) 対象の原著者名・原論文名等のメタな用例が混じる恐れがある。このため、「雑誌名」(当該論文の) 「著者名」等のタグを施したデータコーディングを行い、タグを意識した検索が可能にしてある。

「万葉」を検索、但し 雑誌「万葉」(「萬葉」) を除く。

SJS : Study in Japanese Linguistics Survey DB search

mtoyo@aa.tufs.ac.jp, 2000

『国語学』:「展望」記事検索

国語学会データベース委員会
最新更新日付 2000/11/27

これは、『国語学』の「展望」のうち、著者許諾済みのものについて、ダウンロード用の記事と記事検索を公開するものです。

権利関係

1. この『国語学』「展望」データベースに含まれる「展望」記事の著作権は、それぞれの著者に属します。
2. それぞれの記事本文を、この「展望」データベースによって検索し、検索結果を私的に複製する事、又はその全文(画像を含む)をダウンロードして私的に複製することは、それぞれの記事の著者によって許諾されています。つまり、これらの記事は、自由に検索・私的複製して構いませんが、あくまで私的複製の範囲で御利用下さい。
3. それぞれの記事本文を、この「展望」データベースによってデータベース化し、wwwサーバに置いて一般の検索・ダウンロードに供する事は、それぞれの記事の著者によって国語学会データベース委員会に対して許諾されています。
4. この『国語学』「展望」データベースのデータベース著作権は、国語学会データベース委員会にあります。

検索文字列(正規表現可) [万葉]

検索例
空白で区切ると AND、| (縦棒)で OR、! (感嘆符)で否定

[万葉]葉　仮名|漢字　→　「万葉」又は「萬葉」且つ「仮名」又は「漢字」を含む記事を検索
[平上去入]声　→　「平声」「上声」「去声」「入声」を検索
(状態|作用)性　→　「状態性」又は「作用性」を含む記事を検索
状態|作用性　→　「状態」又は「作用性」を含む記事を検索
接(頭|尾)?[辞語]　→　「接頭辞」「接尾辞」「接頭語」「接尾語」「接辞」「接語」を検索

[万葉]葉　!集　→　「万葉」又は「萬葉」で、「集」は含まない記事を検索

詳細指定
タグ名　検索語(正規表現可、空白で区切ると AND、| (縦棒)で OR), ! (感嘆符)で否定
著者 <author>
題名 <title>
掲載誌名 <serial> ![萬万]葉
例 : ![万葉]葉　→　雑誌「万葉」ではない記事を検索
年紀 <date>
出版元 <publisher>

100 件まで結果を表示 (0 で無制限)

検索開始　やり直し

- 最初のページへ
- 趣意書へ
- 国語学会ホームページへ
- 『国語学』総目次・著者索引へ

国語学会データベース委員会
連絡先　国語学会事務室 〒113-0033 東京都文京区本郷1-13-7 日吉ハイツ404号
電子メール jimusitu@blue.ocn.ne.jp
委員長　山口佳紀

検索結果

SJS search result for `text(万葉), not_serial([萬万]葉)'

2001/10/23:10:11:17 generated by `SJS-frnt.cgi', do not edit

SJS 検索結果 `text(万葉), not_serial([萬万]葉)'

全 11 例中 11 例表示

1. 097-11 文字・表記 (野村雅昭)[原本画像(PDF)] 「国語学」97(1974-06-30) 万葉

   犬飼隆の「万葉“仮名”から仮名へ」(国語学93号、昭48・6)は、その題目から予想されるような、かな字体の起源に関するものでなく、「system としての万葉仮名が内包していたところの、それを母体として片仮名・平仮名が生まれるべき必然性を叙述」したものである。すなわち、ここで述べられるのは、万葉がなに内包される。語形表示の明確さへの要求と、表現・伝達に致命的な支障がおきない範囲での簡略化への指向という二つのベクトルのうち、識字層の底辺への拡大という現象の中で、後者がしだいに強くなり、漢字本来の機能についての「価値ある忘却」がおこなわれ、しだいに「漢字離れして」いく過程である。

2. 129-03 文法(史的研究—古代) (山口佳紀)[原本画像(PDF)] 「国語学」129(1982-06-30) 万葉

   [129-18上]その他、内田賢徳「動詞重複形態の述語」(『帝塚山学院大学日本文学研究』11、昭55・2)、信太知子「上代語における連体形準体法について-- 万葉集を中心にク語法との関連など--」(『馬淵和夫博士退官記念国語学論集』大修館書店、昭56・7)、山口佳紀「形容動詞の成立」(『国語と国文学』58--5、昭56・5)、同「タリ型形容動詞の成立」(『国語国文』50--2、昭56・12)などがあった。

3. 129-03 文法(史的研究—古代) (山口佳紀)[原本画像(PDF)] 「国語学」129(1982-06-30) 万葉

   [129-18上]その他、内田賢徳「動詞重複形態の述語」(『帝塚山学院大学日本文学研究』11、昭55・2)、信太知子「上代語における連体形準体法について-- 万葉集を中心にク語法との関連など--」(『馬淵和夫博士退官記念国語学論集』大修館書店、昭56・7)、山口佳紀「形容動詞の成立」(『国語と国文学』58--5、昭56・5)、同「タリ型形容動詞の成立」(『国語国文』50--2、昭56・12)などがあった。

4. 129-03 文法(史的研究—古代) (山口佳紀)[原本画像(PDF)] 「国語学」129(1982-06-30) 万葉

   [129-18上]その他、内田賢徳「動詞重複形態の述語」(『帝塚山学院大学日本文学研究』11、昭55・2)、信太知子「上代語における連体形準体法について-- 万葉集を中心にク語法との関連など--」(『馬淵和夫博士退官記念国語学論集』大修館書店、昭56

   は、阪倉篤義氏が万葉集の「ゐる」と「をり」の用法に基づいて、前者が「存在の客観的記述」、後者が「存在の主体的描写」を表すとしたのを承け、この区別が抄物の「ゐる」と「をり」についても有効であることを主張している。近藤泰弘「丁寧語のアスペクト的性格--中古語の「はべり」を中心に--」(『辻村記念』平成4)は、さまざまな「はべり」の用法の中に、「あり」や「り・たり」の変容としての状態性の意味を持つものと、アスペクト的には中立で、敬語的な意味を与えるためだけに動詞等に付加されるものの二種類があることを実証した。TAKEUCHI,Lone “ Long-term developments in the Japanese aspecttense system:a case of linguistic drift? ” (Acta Orientalia,School of Oriental and African Studies,London,d1993)は、広範囲の現象に目配りの効いたテンス・アスペクト史の概観になっている。

11. 185-14 数理的研究 (豊島正之)[原本画像(PDF)] 「国語学」185(1996-06-29) 万葉

    [加藤95]加藤浩司(1995)古代語形容詞ク活・シク活考-- 万葉集和歌における用法上の差異に着目して--/名古屋大学国語国文学 76(九五-一一七)

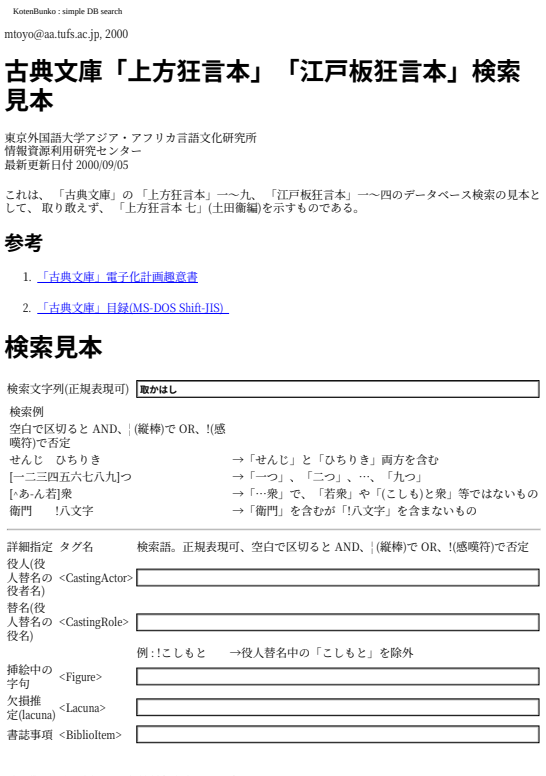
SJS DB project

1.2.2　構造記述が邪魔な実例

原本自体に構造無視がある場合は、構造記述をスキップした検索が要求される。
この極く単純な例は、版本の版心・頁付け (ノンブル) 等である。これらは、原本の構造記述自体から除外するの

が常識的であるが、これ以外に、原本に挿絵・そのキャプション等が割込む場合を例に掲げる。

- **古典文庫・上方狂言本** DB `http://irc.aa.tufs.ac.jp/KotenBunko/`
  AA 研情報資源利用研究センターが、古典文献の翻刻からのデータベース作成実験として、著者全てからの了承の下、実験構築中のデータベース。
  原本 3 オが「取」〜4 ウ「かはした事」と続き、その間の 3 ウ・4 オに挿絵が入る。翻刻本もこの分断を踏襲している。挿絵・キャプションをまたいで「取かはし」を検索したい。



KotenBunko : simple DB search

mtoyo@aa.tufs.ac.jp, 2000

古典文庫「上方狂言本」「江戸板狂言本」検索見本

東京外国語大学アジア・アフリカ言語文化研究所
情報資源利用研究センター
最新更新日付 2000/09/05

これは、「古典文庫」の「上方狂言本」一〜九、「江戸板狂言本」一〜四のデータベース検索の見本として、取り敢えず、「上方狂言本 七」(土田衛編)を示すものである。

参考

1. 「古典文庫」電子化計画趣意書
2. 「古典文庫」目録(MS-DOS Shift-JIS)

検索見本

検索文字列(正規表現可) 取かはし

検索例
空白で区切ると AND、| (縦棒)で OR、!(感嘆符)で否定

せんじ　ひちりき	→「せんじ」と「ひちりき」両方を含む
[一二三四五六七八九]つ	→「一つ」、「二つ」、…、「九つ」
[^あ-ん若]衆	→「…衆」で、「若衆」や「(こしも)と衆」等ではないもの
衛門　!八文字	→「衛門」を含むが「!八文字」を含まないもの

詳細指定	タグ名	検索語。正規表現可、空白で区切ると AND、| (縦棒)で OR、!(感嘆符)で否定
役人(役人替名の役者名)	<CastingActor>	
替名(役人替名の役名)	<CastingRole>	
		例 : !こしもと　→役人替名中の「こしもと」を除外
挿絵中の字句	<Figure>	
欠損推定(lacuna)	<Lacuna>	
書誌事項	<BiblioItem>	

http://irc.aa.tufs.ac.jp/KotenBunko/ (1/2) [01/10/23 11:09:47]

KotenBunko : simple DB search

100 件まで結果を表示 (0 で無制限)

検索開始 やり直し

権利関係

「上方狂言本 七」(古典文庫 409、土田衛 1980)の電子化・翻刻・複製・公衆送信・送信可能化については、古典文庫 主人吉田幸一博士、著者土田衛様よりの御許諾を頂戴している。

東京外国語大学アジア・アフリカ言語文化研究所情報資源利用研究 センターは、学術目的での個人利用に限って、複製を許諾する。

本文ファイルのダウンロード

409-1.lzh 古典文庫409 上方狂言本 七(土田衛) (2000/09/03更新)

- 本文ファイルのデータ形式に就て(2000/09/05)

豊島正之 / mtoyo@aa.tufs.ac.jp

KotenBunko search result for `text(取かはし)'

2001/10/23:11:13:37 generated by `KBD-s.cgi', do not edit

KBD 検索結果 `text(取かはし)'

1 例表示

1. 409 上方狂言本七 (土田,衛)[PDF] 「古典文庫」409(1980,昭和55年10月20日) 128((四ウ)) 取かはし

   かくて,さだの左衛門やしきには,むこ主ぜんのかからう,とね川与一左衛門は,大平七郎右衛門にむかい,「此間お心づかいにあづかり忝し,則今日主ぜん国へかへれば供を申,おいとま申」と云ば,七郎右衛門聞,「□/それはざんねんに存る,扨御とうりうの間に申合し義,かまいてさうい有べからず」と云ば,「何が扨,たがいに,しんもんを取かはした事,何時てもけさう△△△△△付ん」と△△△むこ君花□/がき主ぜん立出給へば,

Koten Bunko project : (temporary mailto mtoyo@aa.tufs.ac.jp)

http://irc.aa.tufs.ac.jp/KotenBunko/KBD-s.cgi [01/10/23 11:13:38]

http://irc.aa.tufs.ac.jp/KotenBunko/ (2/2) [01/10/23 11:09:47]

この部分の XML **ソース**

```
< PAGENOTE page="126"/>
へぞ上りける. < lf/>
< lf/>
かくて, さだの左衛門やしきには. むこ主ぜんのからう. とね川与一
左衛門は, 大平七郎右衛門にむかい.「此間お心づかいにあづかり忝
し. 則今日主ぜん国へかへれば供を申. おいとま申」と云ば, 七郎右
衛門聞.「< Lacuna >□/そ</Lacuna >れはざんねんに存る. 扨御とうりうの間に申合し義. か
まいてさうい有べからず」と云ば.「何が扨, たがいに, しんもんを
取< FOLIA fol="(三オ)"/>< lfPreFigure/>

< Figure >
挿絵< tab/>第一図< origFOLIA fol="(三ウ)"/>< lf/>
挿絵< tab/>第二図< origFOLIA fol="(四オ)"/>< lf/>

< PAGENOTE page="127"/>
第一図 (上段) < lf/>
『みうら介六, 松へ上り』< lf/>
『やつこ八平, 馬引ゐる』< lf/>
『きつね共, さうれいする所』< lf/>
(下段) < lf/>
『こしもとおしげ』< lf/>
『井の上新三郎, かけ付る』< lf/>
『はるひめ, おそれ入給ふ』< lf/>
『かづさがいもとおきく』< lf/>
『ざとう久いち, にぐる』< lf/>
『きつね, 介六< atRight >ニ</atRight >ばけ来る』< FOLIA fol="(三ウ)"/>< lf/>
< lf/>
第二図 (上段) < lf/>
『井上新三郎』< lf/>
『大殿さだの左衛門』< lf/>
『若殿いおり』< lf/>
『むこ主ぜん殿』< lf/>
『かづさが女房おすが』< lf/>
『大平七郎右衛門』< lf/>
『主ぜんからう与一左衛門』< lf/>
(下段) < lf/>
『左衛門, りつふく』< lf/>
『七郎右衛門, うつてに行んと云』< lf/>
『新三郎, せうこに立』< lf/>
『おすが, おしとめる』< lf/>
『かづさいもとおきく』< lf/>
『かづさ, せんぎを聞』< lf/>

< PAGENOTE page="128"/>
(上段) < lf/>
『かづさの介九郎』< FOLIA fol="(四オ)"/>< lf/>
</Figure >
< lf/>
かはした事. 何時てもけさう< Lacunae >△△△△△</Lacunae >付ん」と< Lacunae >△△△</Lacunae >むこ君花< Lacuna
>□/が</Lacuna >き
主ぜん立出給へば. < lf/>
若殿いおり出, おいとまごいの給ふ. < lf/>
所へ, 大殿左衛門出給ひ.「主ぜんかへり給< Lacuna >□/ふ</Lacuna >か, 此度, 身かきしよ
くに付見まひと有て. はる〃〃参らるゝだん, まんぞくに存る. 本ぶ
くいたしたれば, おつ付はる姫をおくるでござらふ程に. ふびんをく
はへて下され.」「其だんはお心やすう思召ませ.」「おゝうれしう存
る. むこ殿おたちじや. かづさふうふはいづくにゐるぞ」との給ふ所
へ. < lf/>
```

(**古典文庫** 409 **土田衛編「上方狂言本七」より**)

2 XML ドキュメントの交換可能性

2.1 「互換性」が確保する「互換」とは何か

SGML/XML で文書の「互換」を確保する為には、次の様なものが試みられて来た。

1. 同一 DTD を使う
   → 拡張性皆無で、実際には実施不可能。
   同一 DTD が強制出来る程に統制が取れた文書作成集団なら、そもそも「互換性」等を気にする必要は無いのではなかろうか
2. 極力構造性の無い DTD にし、共通辞書 (「ボキャブラリ」) 又はタグ名対応辞書を作る
   → ebXML 等、「業界標準」作成の模索で使われた方式。共通辞書が作れる位なら、文書標準位、軽いものだろう。
3. DTD をパーツ化し、タグ名は決め打ち、部分詳細化は parameter entity で行う。

→ 一見、拡張性がある。TEI を始め、過去の多くの SGML 大アプリケーションが取った方法であるが、SGML/XML の文法的制約、特に名前スコープの無さから、重厚長大複雑怪奇な DTD になり、メンテナンスが困難なばかりか、名前の衝突など、ユーザには解決困難な問題を引き起こす。しかも、タグ名を決め打ちした処で、タグの表現内容に就ての合意が無ければ無意味である。これは TEI の破綻の最大の要因となった。

4. DTD とは別のスキーマを設け、タグの構造的意味付与を、スキーマ文法記述で行う。
   → XML スキーマの取る方法。スキーマ自体が乱立の状況で、安定して使えない。又、タグへの構造意味付与も、表現内容の合意が無ければ無意味である点は同じ。

2.2 attribute の付与する意味は、交換可能ではない

attribute は意味付与以外の機能を持たないから、element 名よりも、遥かに意味に直結している。(意味に直結していなければ、attribute を振る意義がそもそも無い)。

2.2.1 attribute の詳細化と交換可能性は相容れない

attribute は、詳細化すればする程、交換不能になる宿命を持つ。DTD の詳細化は共通 DTD 一つであるのに、実際の文書が要求する詳細化の枠組みは、文書ごとに異なるからである。

例えば、「日本電子出版協会電子出版交換フォーマット」JepaX (`http://x.jepa.or.jp/jepax/`) は、element による書籍の構造記述は殆ど放棄して <`div`> 一辺倒とし、その代りに <`div`> に attribute を振って意味付与しようとして、attribute に下記を定義している。

> ”表紙”、”あらすじ”、”抄録”、”著者紹介”、”とびら”、”謝辞”、”献辞”、”序文”、”凡例”、”目次”、”図版目次”、”部”、”篇”、”章”、”節”、”項 1”、” 項 2”、” 項 3” 、”項 4”、”項 5”、”項 6” ”あとがき”、”解説”、”付録”、”索引”、”用語集”、”年表”、”関連書籍”、”奥付” ”引用”、”詩”、”プログラムリスト” 、”参考文献”、”囲み” …

「将来の仕様更新時に追加されていく可能性がある」と明記されている (確かに「項 *n*」はどんどん追加されそうである) が、これだけでも「序文に献辞が含まれていたら ?」「参考文献と関連書籍の差は ?」「項 1 の下に更に項目列挙があったら ?」「引用がプログラムリストだったら ?」「プログラムリストが詩だったら ?」[3]等々、実際のコーディングでは様々な疑問が湧いて来る。送り手のコーダ当人に疑問が湧く様では、受け手はとても安心して受け取れない。

2.2.2 attribute のスコープの問題

attribute は、element をスコープに持つ。(これが、SGML では唯一の名前スコープであった)。従って、global attribute というものが存在しない。

TEI(`http://www.tei-c.org/`) は、この解決のために、全ての attribute の後ろに entity reference で global attribute を埋め込むという挙に出たが、これは徒労である。何故なら、異なる element の attribute の意味が同じである保証が全く無いからである。attribute は element local なので、異種の element に同じ attribute を振っても、それらの属性が同一である事の表現にはならない。

<`thesis type="manuscript"`><`book type="manuscript"`>

等と振る事自体は勿論自由だが、SGML/XML の仕様からすれば、これは振られた attribute がたまたま同一の manuscript という文字列だったというだけの事で、両者にそれ以上の如何なる関係も推定出来ないから、勿論両者の同定も出来ない。実際

<`temparature status="high"`><`tide status="high"`><`quality status="high"`>

を同定されては困るだろう。

この様な attribute の仕様は、同一 attribute を inherit しながら順次詳細化して行く polymorphism が、element 間で行う事が出来ない事を意味する[4]。これは、電子文書のオブジェクト化には全く反する性質である。XML の namespace

[3] 世の中には、perl で書いた詩 (Larry Wall 作) というものも存在する。

[4] 尚、TEI-P3 は、こうしたオブジェクト指向化を element class という概念を立てて行なっているが、実はそこで行なっている inheritance は、上位の attribute 記述を手作業で (!) コピーして来る (3.7.1) といった体のものである。これも、TEI のせいというよりは、SGML の記述力不足というべきであろう。

仕様は、この様な local 仕様を element にも適用しようとするものである。

3 交換可能性の無さは XML のせいではない

こうした XML 文書の意味の面での互換性の無さは、そもそも XML の問題ではない。

タグによる情報交換のためにはタグの指示する概念に就ての共通理解が必要であるが、それが存在しないのが、真の問題である。

TEI 書誌記述を例に取れば、<docTitle> というタグは、TEI 仕様書は形式を <titlePart main="...", sub="...", desc[ription]="..."> と規定するだけで、title とは外題か内題か、外題にしても題簽か扉か、2 冊目のみに原題簽が残る時の他の冊の打付け書きとの区別はどうするか、角書は含むか、等の基本的な書誌事項の規定が何も無い。つまり、<title> タグ概念の詳細に合意が取れないのであるが、これは、TEI の不備というより、そもそも書誌学に於て言語・時代・分野を超えての「title」という共通理解が存在しないからに他ならない。(言語・時代・分野を超えなければ、TEI の存在意義はそもそも無い)。つまり、情報交換規格の不備ではなく、書誌記述に対する共通理解形成の不備であり、書誌学が、実は交換可能な概念を持っていない事が、ここに露呈しているのである。

事は書誌学に限らない。HTML の <p> で頻用される「paragraph」自体、定義は困難で、W3C の HTML specification (4.1) http://www.w3.org/TR/html4/struct/text.html#edef-P は、「Authors traditionally divide their thoughts and arguments into sequences of paragraphs. 」と tradition に責任を押し付けている。しかし、これは正しい態度であって、早く檜山 (1998) の指摘がある。「段落や見出しは、… 文書の記述様式である。国語教育や読み書きの経験で習得され、社会や特定集団内で共有される規範である。この規範を写し取るには、目的の文書文化 (読み書きの習慣と形式の総体) のなかに身を置き、経験するしかないだろう。」

こうした「文化的に定義される術語」という側面以外に、研究対象として見た場合、title や author などの曖昧な用語は、曖昧であるからこそ、概念のバスケットとして使えるのであり、研究を進展させるものである。(例えば title page の研究 Smith(2000) を title の厳密定義から始める訳には行かない)。「short-title catalogue」と断った目録が編まれるのは、title に一貫した特徴が無いからである。交換可能な程に精密化された概念を求めたのでは、研究の余地が無い。title や author といった概念語は、記述表現 (「これが title である」) ではなく、encoding を行った者の解釈表現 (「私はこれを title と呼ぼう」「title に相当するものとして、私はこれを挙げる」) であり、一つ一つの用例に則して解釈を理解するしかなく、そもそも「交換」には向いていないのである。こうした意味からも、文書の構造を出来るだけ精密に定義して構造的意味によって文献を記述するという発想は、全く文献学的研究向きではないと言えよう。

以上から、文書構造を SGML/XML 風の hierarchy で記述する事には、文献学的研究に取っては何の意味も見出せない。少なくとも、文書構造自体の「交換」が必要な局面も、有用な局面も存在しないと考える。TEI の様な規準が役に立つとしても、せいぜい (作業者間で合意を見ているタグ付け規約での) タグ名・attribute 名統一等、コーディング担当者のマニュアル・心覚え程度であろう。

実際、TEI comformant なドキュメントから、<title> タグを拾う事で、何が主張出来るだろうか？ 抽出したデータから書誌目録が編纂出来るだろうか？実際は、当該文書のタイトルに加え、関連文書のタイトル、その文書の注釈の為に引用された書・用語の解説書のタイトルまでも <title> であるので、<title> で収集されるのが玉石混交であるだけでなく、本来拾いたいものが、<title> でマークされずに取り落されている可能性もある。

4 検索に望まれる技術

XML のマークアップ校正を繰り返したテキストに対して、tag を剥ぎ取った形での頒布を求められるのは、大変空しいが、しばしばある事である。これは、XML データの検索技術等、(今の処) 誰も信用していないからであろう。

4.1 お仕着せ検索の問題

Langland, William(ed. Adams, Robert, et al (2000)) は、中世英語彩色写本を「TEI-P3 full comformant」でマークアップし、ビュア (browser) 兼検索ソフトウェアを添えたものであるが、その Readme は、次の様に述べて、このソフトウェアを実質見限っている。

> However, since the browser does not permit full display of every feature of the manuscript that we have recorded in SGML markup, you may find it useful for some purposes to search the *.sgm text files with an ASCII editor rather than search with the browser. ...
> Readers should, therefore, expect the search facility on the browser sometimes to be less thorough and useful than that on a word processor or text editor operating on the plain ASCII text.

TEI-P3 full comformant なのはマークアップであって、ソフトウェアではないのである。TEI-P3 を活かした汎用の検索ソフトウェアが如何に難事かを伺う事が出来よう。

4.2 手作業の現状・望まれる技術

(1) well-formedness の検証・エディタ

XML エディタ数種を試したが、TagEditor (アンテナハウス `http://www.antenna.co.jp/`) に落ち着いた。(因みに、試した中で最も安価であった)。

a) well-formed でない文書、作成中の文書を読ませても、最後まで読む。
呆れた事に、世の中には「well-formed ではない」とエラーメッセージを出して読み込みを中止する XML エディタも、売り物になっている。これでは、既存の文書を読み込んで XML 化する等は、不可能である。ゼロから XML 文書を手で書く以外には使えないエディタというデザインのセンスは、XML の利用局面を全く理解していないと思う。

b) 文書の encoding 指定を無視して読み込める。
文書が自分自身の encoding を指定するのは、クレータのパラド クスであって無意味である[5]。これは、特に XML 文書が SMTP 等で転送された時に然りであるが、こうした「自分自身を裏切る」指定をした文書を読ませると、「encoding が違う」と言って止まる御粗末なエディタすら売り物になっている。

c) 文書を排他的に開かないで、他からのアクセスも許す。
XML 文書の宿命として一旦全文を slurp せざるを得ないが、排他的に開き続ける事がないので、別のソフトウェアで更新を掛ける事も出来る。他からの更新が掛かると、TagEditor はそれに気付き、読み直すか否かを問い合わせて来る。slurp 後も開き続け、他からの読み込みすら sharing violation を引き起こすエディタもある。

(2) 検索

上述の AA 研のデータベースでは、perl 経由の DOM (Level 1, `http://www.w3.org/DOM/`) で各階層のオブジェクトを取り出し、それを階層ごとに結合したものを perl の DBI (Database interface) 経由で SQL 系 DBMS (mySQL) のテーブルに格納し、これを同じ DBMS で検索・提供している。これだと一旦「下準備」が済めば高速であるが、タグの種類が多くなって来ると「下準備」に手間が掛るし、DOM の利用も実質的には各データベース毎に補助的なルーチンを補わねばならず、手作業に近い。ここを一般化しようとすると、SQL のテーブ

[5] 豊島 (1994) 参照。

ル定義から可変にしなくてはならず、これは perl の DBI・java の JDBC 経由等で可能であるが、余りに大事になるので諦めた。尚、XSLT は、構文が美しくない事[6]、動作に今だ不審な点が残る事等から、利用していない。オブジェクトの階層を意識した検索が、もう少し簡単に出来ると有り難い。

期待する技術としては、on-the-fly で DOM を意識したフィールドを取り出して検索出来る事が第一である。これ自体は perl や java で可能だが、その場合、インデクシングや検索自体まで perl や java で書く破目になる。これは避けたいので、効率の良い格納・検索等の処理は、背後に DBMS が位置して処理して欲しい。現在、こうした関係を綺麗に解決して呉れるプロダクトを探している処である。例えば、SQL 系で、DOM による view を可能にするか、select 句に DOM 呼び出しを書いたかの様に振る舞うプロシージャがあれば、かなり楽な筈であるが、実例には、未だお目に掛かれない。

XML native Database には大変興味があるが、今の処、とても文系の一研究所が試しに購入出来る価格ではないのは残念である。

5 引用文献

1. Bourret,Ronald(2000) XML Namespaces FAQ (`http://www.rpbourret.com/xml/NamespacesFAQ.htm`)
2. Langland, William(ed. Adams, Robert, et al (2000)) Corpus Christi College, Oxford MS 201 (The Piers Plowman Electronic Archive, Vol.1, The University of Michigan Press, [CD-ROM], ISBN 0-472-00275-9)
3. Smith, Margaret(2000) The title page – its early development 1460-1510 (Oak Knoll press)
4. 出雲朝子・豊島正之 (1993) 『玉塵抄』と計算機 II (文部省科学研究費研究補助金「『玉塵抄』の計算機処理に就ての発展的研究」研究成果報告書)
5. 豊島正之 (1992) TEI から見た SGML のはなし (情報処理語学文学研究会会報 12、`http://jcs.aa.tufs.ac.jp/mtoyo/TEI/JALLC-12-TEI.pdf`)
6. 豊島正之 (1994) TEI-P3 から見た SGML のはなし (情報処理語学文学研究会会報 15、`http://jcs.aa.tufs.ac.jp/mtoyo/TEI/JALLC-TEIP3.pdf`)
7. 檜山正幸 (1998) 文書型定義 (DTD) とその設計 (`http://www.saiensu.co.jp/ct_Fresource/199809/On-DTD.htm`)

[6]その必然性も無いのに、無理やり XML で記述する事にしてあるからであろう。