人文科学とコンピュータ　55－7
（２００２．７．２６）

日本古典文学本文データベース再構築における諸問題
－XML化の現状と課題－

安野一之
文部科学省大学共同利用機関　国文学研究資料館

日本古典文学本文データベースは旧版『日本古典文学大系』（岩波書店 1957.5～1963.4）全100巻、約600作品を収録したフルテキストデータベースである。本データベースは古典文学作品の本格的なデータベースとして、これまで高い評価を得てきた。しかし、基本設計から15年近くを経過し、様々な意味で再構築が要請されてきた。これまでKOKINルールと名付けられた特定マークアップ言語で記述されていたデータを、XMLに変換する作業はその一環である。本発表ではデータベース再構築の現場から、具体的にどのような問題が発生し、どのように対応していったのかを報告したい。

Examination of various problems in Japanese classical literature data base restructuring.

Kazuyuki YASUNO

(National Institute of Japanese Literature)

The full-text database of Japanese classical literature which collected “Nihon koten bungaku taikei” (IWANAMISHOTEN 1957.5-1963.4). Up to now, this database has obtained the high estimation as a real database of the classics literary work. However, about 15 years pass from a basic design, and the necessity restructured with various meanings has come out. Work to convert the data described by the specific standard generalized markup language named KOKIN rule into XML. It reports what problem occurred concretely by the process of the data base restructuring in this thesis.

はじめに

現在、国文学研究資料館では様々なデータベースが運用されている。日本古典文学本文データベースはその一つであり、名前からも分かるように、テキストデータベースの一つである。本データベースには岩波書店から刊行されていた旧版『日本古典文学大系』(1957.5～1963.4)全 100 巻、約 600 作品全てが納められている。

『日本古典文学大系』に収録された作品は、古事記から浮世草紙まで多種多様であり、使われている文字種、表記法は膨大な数に上る。日本古典文学本文データベースはそれら全てをマークアップし、電子テキスト化するにあたって独自のマークアップ言語を定義した。これを KOKIN ルール (KOKubungaku Information)と呼んでいる。KOKIN ルールは基本的には SGML と共通する性質を持つが、古典文学テキスト用に特化したマークアップ言語であり、汎用性が高いとは言い難い。しかし、実際の古典作品から導かれた簡潔で明快な構造は、本データベースの主な利用者である古典文学研究者から高い評価を得、KOKIN ルールを利用したデータベースも構築されている。

XML 化への道程

KOKIN ルールは日本古典文学という、これまで資料の電子化が困難であった分野における嚆矢として高く評価される一方、その基本設計から 15 年近くを経過し、様々な意味で問題が生じてきたことは否めない。

第一の問題としては、KOKIN ルールを設計した時点では、一般的とは言い難かったマークアップ言語が、SGML から XML へと移行する過程で急激に普及してきたことが挙げられる。今後の Web 対応を考えたとき、標準的なマークアップ言語に準拠することは様々な意味でメリットがある。システムレベルで考えれば、Z39.50 などを介して、国内外のデータベースと有機的な連鎖を期待することが出来るし、二次加工を行う際にはツールを独自開発しなくても、安価な市販のツールが用意されている。また、パーソナルレベルで考えた場合も、これから KOKIN ルールを覚えるよりも、XML を覚える方が簡便である。

第二の問題として、これは必ずしも KOKIN ルールの問題ではないのだが、漢字コードの変化を挙げることが出来る。日本古典文学本文データベースの設計段階の文字セットは、大型汎用機上で運用されていたこともあり、JIS X208-1978 であった。しかしその後、ハードウェア環境の更新に伴い、JIS X208-1978 が JIS X208-83 に機械的に置き換えられるなど、いわゆる旧 JIS と新 JIS の混乱が起きてしまった。また、大型汎用機にあわせて設定された外字も、サーバ上では反映させることが出来ず、ハードウェアの更新が却ってデータベースの質を下げてしまうと言うジレンマに陥ったのである。

また、本データベースが試験公開された結果、多くのユーザーから声が寄せられるようになった。古典文学の専門家であるユーザーからのフィードバックは非常に貴重なものであり、データベースの信頼性を高めるためにも、データのエラー箇所には早急に対処することが求められた。

作業の現場から

日本古典文学本文データベースを XML 化するにあたって最初に取り組んだ問題は KOKIN ルールで記述されたデータを変換するプログラムを作成することであった。幸い、KOKIN ルールを SGML に変換する検証実験[1]は既に行われていたので、それをベースに DTD を新たに定義した(図 1)。同時に機能拡張の一環として、『日本古典文学大系』の頭注、校注に関しても新たにデータ化を試みた。そもそも『日本古典文学大系』は出版当時、厳密な校訂、精緻な注釈を施すこ

---

1 原正一郎,安永尚志"国文学研究支援のための SGML/XML データシステム" 情報知識学会誌 Vol.11,No.4, 2002.

とを旨として構想された叢書であり、頭注、校注の入力は古典研究者からの強い要望に応えるものであった。頭注、校注は独自の書式をもつので、本文部分とは多少異なる DTD を設定することになった（図 2)。これらを組み合わせて印刷用 PDF も作成した。

DTD を巡る問題

KOKIN ルールから XML への変換の中はプログラムによって行ったが、いくつかの問題が発生した。そもそも、KOKIN ルールは本の物理的な構造と論理的な構造を一体化させた考え方、すなわち本文の位置情報を忠実に再現することを目的としており、換言すれば『日本古典文学大系』をデジタルアーカイブとしていかに忠実に再現できるかという試みでもあった。しかし、それゆえにページ、行と言った物理構造と、短歌、連歌と言った論理構造が混在する形となり、複雑な構造を持つことになってしまった。特に長歌（万葉集など）のように複数行にわたる和歌の場合、物理構造と論理構造の線引きは困難であり、作品の内容解釈に踏み込む必要が生じた。また、XML では浮動要素（illust,filename,lineno）が扱えないので、出現箇所すべてに定義する必要が生じ、DTD の制約も緩くせざるを得なかった。

次に問題となったのは、『日本古典文学大系』の複雑な傍記情報の扱いであった。厳密な校訂を経て編集されたこれら古典作品には、通常の傍記（右傍記）の他、左傍記、左右傍記、それぞれの二重傍記、割り注等、複雑な文書構造を持っている。これは W3C Ruby Annotation で定義されているルビ規則より遙かに多く、新たなルールを設定する必要が出てきた（図 3)。

また、全作品を分析してから DTD を作成したわけではなかったため、新しい構造が出現するたびに修正作業が必要となり、作業を繁雑なものにした。

漢字を巡る問題

漢字コードの問題は、非常に複雑であり、デリケートな問題でもある。特に古典文学を扱う際には、規範化され得ない文字、外字の問題は不可避であり、本データベースの草創期から様々な試みがなされてきた。国文学研究資料館では本データベース構築時に約 2000 文字の外字セットを作成しているが、先述の通り、ハードウェア環境の更新に伴い、現在は利用されていない。今回の再構築作業ではより柔軟で汎用性のある外字対応が求められた。

現在、汎用性のある外字セットはいくつか存在するが、本データベースの特殊性、国文学者を中心としたユーザーの利便性、将来的にも安定して利用することが可能か否か等々の要素を勘案した結果、「今昔文字鏡」の文字コードを採用することとなった。外字出現箇所には"&m123456"といったコードを打ち込み、随時文字鏡フォントを参照することとした。これにより、これまで外字として表示不能であった文字の大多数を表示することが出来るようになった。これはユーザーのパソコンに「今昔文字鏡」がインストールしてあれば言うに及ばず、Web 経由で文字鏡研究会の GIF リンクサービス（http://www.mojikyo.gr.jp/gif）に接続することによって表示することも可能であり、ユーザーフレンドリーなシステムになったと言えよう。

しかし、本データベースの外字全てが「今昔文字鏡」に包含されるわけではない。全 100 巻の『日本古典文学大系』に含まれる文字数はおよそ 3000 万文字と言われており、「今昔文字鏡」に含まれない文字、言ってみれば「大系文字」が出現する。この場合、出現箇所には"&k123456"というコードを打ち込み、出現履歴を管理している。現在、全 100 巻中の約 1/4 を終えた段階で、「大系文字」は 200 文字程出現している。これらの文字をどのように扱うかはいくつかの方法を現在検討中である。

漢字を巡る問題としては、外字の他にも、字体の包摂規準を巡る問題が起こった。現在、漢字は JIS X 208-1990,Unicode(UCS-2),今昔文字鏡,外字という 4 つのフェーズにわたっているが、JIS の包摂規準では割り切れない微妙な差異をどのように扱うか明確な規準を立てることが出来なかった。データベースの主要機能である検索を重視すれば全ての漢字は JIS X 208 の世界に囲い込む必要があり、原本（古典文学大系）を忠実に再現しようとすれば多くの漢字は今昔文字鏡、外字にならざるを得ない。この判断基準が曖昧なまま作業を進めたため、結果として混乱を招い

てしまった。現在でもこの問題が完全に解決したとは言えないのだが、漢字を決定するプロセスを逐一記録することによって作業者全員が情報を共有し、ポリシーを確立しつつある。今後の課題としては、蓄積された情報を元に明快なマニュアルの策定があげられる。

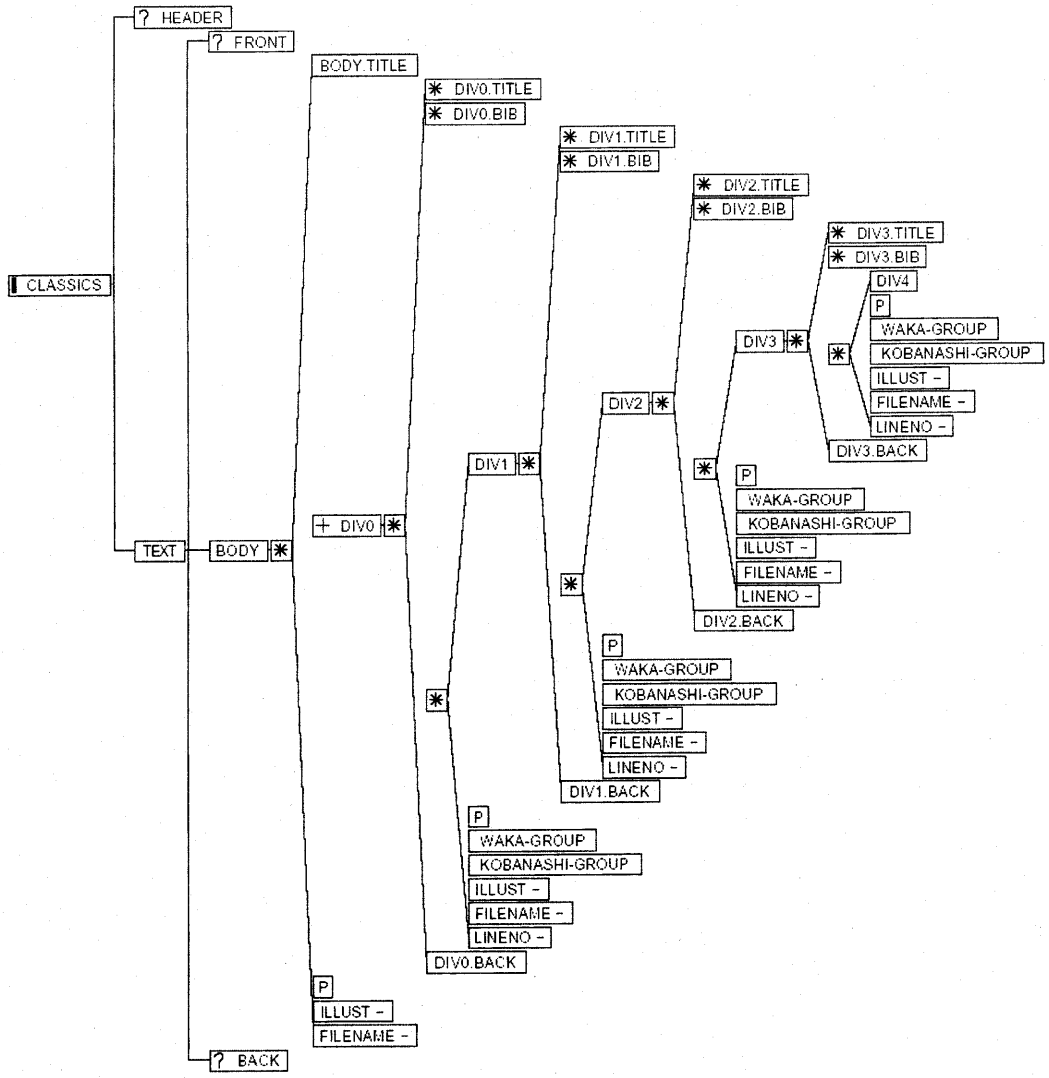
総括

日本古典文学本文データベースは古典文学のフルテキストデータベースとしては最大級のものであり、「マークアップ」という言葉が一般的ではなかった時代にKOKINルールが果たした役割は高く評価されるべきである。しかし、ネットワーク利用が前提となった今日の状況にあっては、より汎用性の高いXMLに変換していくのは必然と言えよう。だが、ルビの問題や、漢文の返り点の扱いなど、現時点では解決できない問題も山積している。今後、XMLが発展していく中で解決されることを期待したい。

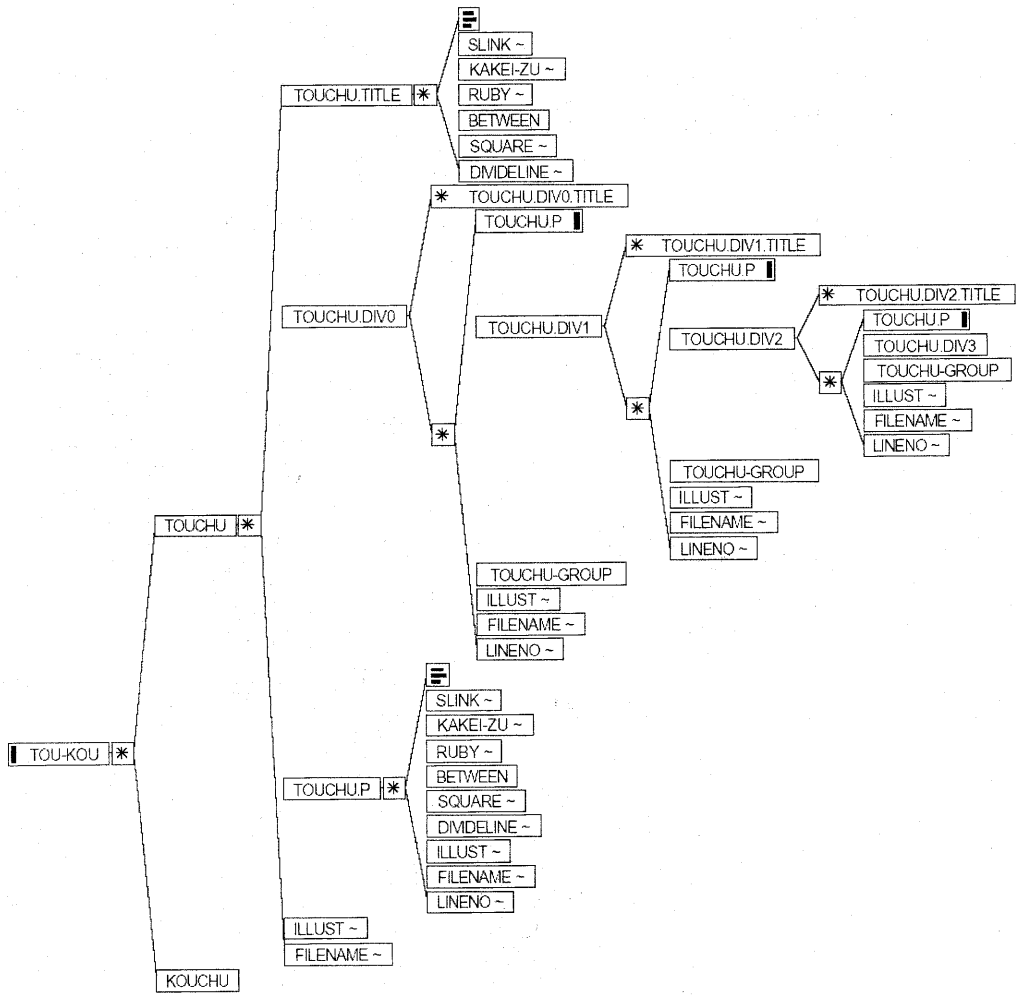
今回の改訂作業の目的は大きく二つに分けられる。一つは特定マークアップ言語から汎用マークアップ言語への変換。そしてもう一つはハードウェア環境の更新に伴って発生した外字の問題、並びに文字コードの問題である。文化の継承を謳う古典作品のデータベースが、数年サイクルのハードウェアの更新に影響を受けるのは由々しき問題であると言えよう。

データベースの更新という作業は日常的なものであり、時間的にも人的にも連続性が求められる。今回、再構築の現場から見た場合、こうした情報をいかに蓄積し、継承していくかという問題の重要性を痛切に感じた。今後は完成されたデータを公開するだけでなく、その作成過程をも含めた記録を公開していく必要があると思われる。特にDTDや外字の問題は、今後行われる各種データベースの更新作業の際にも、必ず問題となることは想像に難くない。これらの情報を集約し、共有できる環境の整備が待たれる。

（図１）　日本古典文学本文データベース本文 DTD



（図２） 日本古典文学データベース頭注 DTD



（図３）　Ruby タグの構造

（１）KOKIN ルールの傍記の種類

boukiso	傍記素（ルビベース）
ruby	本文の右にある傍記データ
lruby	左傍記
lrruby	左右傍記
rruby	左右傍記のうちの右傍記
lruby	左右傍記のうちの左傍記
drruby	右二重傍記
rsideduby	右側傍記
lsideruby	左側傍記
dlruby	左二重傍記
rsideduby	右側傍記
lsideruby	左側傍記
dlrruby	左右二重傍記
drruby	右二重傍記
rsideduby	右側傍記
lsideruby	左側傍記
dlruby	左二重傍記
rsideduby	右側傍記
lsideruby	左側傍記

（２）W3C Ruby Annotaion の要素

ruby	全コンテナとして機能するインライン要素
rbc	ルビベースコンテナ
rtc	ルビテキストコンテナ
rb	ルビベース
rt	ルビテキスト
rp	ルビ括弧

（３）XML でのタグ付けルール案（W3C Ruby Annotation 類似のルール）

＜右傍記＞

```
<boukiso>鳴呼</boukiso><ruby>ああ<ruby>
ああ<lruby>
    ↓
<ruby>
<rb>鳴呼</rb>
<rt>ああ</rt>
</ruby>
```

＜左傍記＞

```
<boukiso>鳴呼</boukiso><lruby>
    ↓
<ruby>
<rb>鳴呼</rb>
<rt position="left" >ああ</rt>
</ruby>
```

＜左右傍記＞

```
<boukiso>鳴呼</boukiso>
<lrruby><rruby>ああ<rruby><lruby>かんどう<lruby></lrruby>
    ↓
<ruby>
<rbc><rb>鳴呼</rb></rbc>
<rtc><rt>ああ</rt></rtc>
<rtc><rt>かんどう</rt></rtc>
```