古典籍のXML化プロセスにおける諸問題

－日本古典文学本文データベース再構築作業を通して－

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

Various Problems on transcription of the Japanese Classical Literature Database.

Kazuyuki YASUNO
(National Institute of Japanese Literature)

In this report, various problems arisen in the course of the conversion processes from the full-text database for Japanese Classic Literatures to XML database will be described.

This database had been described under so-called KOKIN rule, however it is necessary to set up a definite DTD. Furthermore, it will be required to expand W3C Ruby annotation for the expression of the complex ruby, which is specific in the classic literatures for the conversion to XML.

It is not easy to express the Kanji used in classic literatures. In this report, we solved this question by applying "Konjaku Mojikyo".

The remaining problem is how to describe the conversion points of "Kanbun " and how to express the Kanji, which may not be expressed by "Konjaku mojikyo".

1. 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もその一つであり,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作品全てが収録されている.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高い評価を得ており,登録利用者数も 2000 人を越える.しかし,基本設計から 15 年以上を経過し,様々な意味で再構築が要請されている.その一環として,これまでKOKIN (KOKubungaku Information) ルールと名付けられたマークアップ規則で記述されていたデータを,XMLに基づくデータに変換する作業を行っている.

本稿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するとともに,XML データが古典文学研究にどのような影響を与えうるのかを検証した.

2. 現状と課題

KOKIN ルールは国文学研究資料館が独自に規定したマークアップ規則である [1] .国

文学研究資料館ではこれまでに「日本古典文学大系」の他,「噺本大系」,「仮名草紙集成」（東京堂出版）など,約 150 巻,およそ 5000 万文字の電子化を行ってきた.その結果,KOIIN ルールはこれら日本古典文学の幅広いジャンルにわたるテキストの電子化に対応可能であり,古典テキストを電子化する上で必要とされるマークアップ機能を満たしていると見なすことができる.このように実際の古典作品に即して検証された KOKIN ルールは,日本古典文学本文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［2］.

しかし,KOKIN ルールにも幾つかの問題点が存在する.まず,KOKIN ルールは国文学研究資料館が独自に開発したマークアップルールであり,データ処理のためのあらゆるツールを開発する必要があった.また,KOKIN ルールは全作品に共通する文書型のみを考慮していたため,作品ごとに現れる固有の記述や構造に対応する際,同一記号に複数の意味を持たせてしまったなど,ルール構文上の曖昧さを残してしまった.

古典テキストを電子化する上で,避けて通れない問題の一つに漢字処理が挙げられる.「日本古典文学大系」に含まれる文字数は約 3000 万文字と言われており,数多くの外字が含まれる.日本古典文学本文データベース構築時のの文字セットは,大型汎用機（HITAC 860/60）上で運用されていたこともあり,JIS X208-1978 であった.しかし,国文学研究資料館の情報システムが,Unix サーバを基本とする分散管理システムに変更されたため,文字コード系が JIS X208-83 となり,いわゆる旧 JIS と新 JIS の混乱が生じた.また,大型汎用機にあわせて作成された 2000 字余りの外字フォントも,新しいサーバ上では利用することが困難である.情報システム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事態に陥ったのである.

こうした状況を踏まえ,日本古典文学本文データベースの発展的運用を考えたとき,データの可搬性と可読性が保証された汎用マークアップの適用が必要不可欠であると考えられた.XML はそのような汎用マークアップ言語の一つである.予備実験により KOKIN ルールは XML DTD で記述可能であることが確かめられた.また DTD を定義することは,古典本文の構造を明確化することにつながる.加えて Web への対応が容易であること,市販の様々なツール類を利用できることなども,XML を導入した理由である.一方,国文学研究資料館で開発が進められているコラボレーションシステムでは,データの共有化がテーマであり,全データの XML 化が前提となっている.古典本文の XML 化が進めば,本データベースも他のデータベースとの共用が可能となる.

また,今回の再構築作業では,これまでに発見された誤字,脱字等のエラーの修正,原本と照らし合わせた厳密な校閲作業を行い,信頼性の高いデータの構築に努めた.

3. XML 化への道程

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［3］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.しかし,対象とするテキストの量が多く,なおかつ多様性に富んでいたため網羅的な DTD を作成することが困難であった.そこで KOKIN ルールと同様

にテキストの基本的な論理構造（タイトル,章など）,レイアウト構造（ページ,行など）および傍記などの注釈のマークアップを対象としたDTDを設定した.

同時に,一部の作品の頭注,校注に関しても新たに電子化を試みた.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とを旨として構想された叢書であり,刊行から 40 年以上経過した現在でも,頭注,校注情報は研究者に参照され続けている.これらは本文部分とは異なった独自の書式を持つので,新たに DTD を定義することになった.また,これらを組み合わせ印刷用 PDF も作成した.

3-1. DTDを巡る問題

SGML と異なり XML では浮動要素が定義できない.そのため,絵や割注などの要素を記述するDTDは制約を緩くせざるを得なかった.つまり定義したDTDに基づいたパーサは,絵や割注などについての正しいマークアップ以外のマークアップを許容してしまう可能性がある.同様に,SGML/XMLで前提となっている論理要素の「入れ子構造」に合致しない部分があった.この部分については,KOKINデータを訂正し,しかる後にXMLに変換する必要が生じた.

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.W3C Ruby Annotation で定義されているルビ規則では対応しきれなかったので定義を拡張する必要があった.

漢文の返り点も傍記情報の一種ではあるが,通常の傍記とは異なり傍記のターゲットとなる文字がなく,文字と文字の間に位置するものである.また,返り点は中国語文法の漢文の文字列を日本語文法に合わせて転倒させるものであり,文章の構造を崩してしまう.これらの理由により,今回の作業では漢文部分のマークアップを諦めざるを得なかった.しかし,古典文学における漢文の持つ意味は極めて大きく,今後,何らかの対応をしていく必要がある.

3-2. 漢字を巡る問題

漢字の問題は,古典文学を扱う際にはそれ自体が研究対象になるため,慎重に向き合わざるを得ない.既に述べたように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たが,現在で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,広く使われている外字セットはいくつか存在するが,古典作品固有の文字が頻出する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「今昔文字鏡」を表す"m"と文字鏡番号である6桁の数字を組み合わせ,"&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る.

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.「日本古典文学大系」固有の文字も数多く含まれており,全体の約半数を終えた段階で 300 文字程出現している.これらの文字に対しては,出現箇所に古典文学大系を意味する"k"と 6 桁の数字を組み合わせ,"&k123456"というコードを打ち込み,出現履歴を管理している.「今昔文字鏡」に含まれる文字と異なり,これらの文字は現段階では表示することが出来ない.今後どのように扱うかはいくつかの方法を検討中である.

4. おわりに

日本古典文学本文データベースを XML 化することは,KOKIN ルールというローカルなルールで記述されたデータの汎用性を高め,Web サービスへの対応など様々な機能拡張の可能性をもたらしてくれた.

今回の作業はまた,これまで KOKIN ルールでは許容されてしまってきたデータ記述の曖昧さを明確化する作業でもあった.その中で完成した DTD は今後,古典文学作品を電子化する上で重要な意味を持ちうると考えられる.そしてまた,XML 化がもたらした明確化は KOKIN ルールのみならず,原本である「岩波古典文学大系」の記述の曖昧さをも浮き彫りにした.特に語り物と呼ばれるジャンルの作品ではテキストそのままでは理解困難な箇所も出現し,口承文芸のテキスト化の困難さを物語ると同時に,テキストとしての不完全さを知らせてくれることにもなった.

今回の作業でもっとも多かったのが漢字を巡る問題であった.これについては稿を改めて論じる必要があるが,今回の作業で心がけたのは,文字を選定することだけではなく,そのプロセスをいかに記録に残すかということであった.それ故,漢字を選定する際には候補となった文字も記録し,どのような経緯で最終決定したのかを記録していった.

現在,全体のほぼ半分の XML 化を完了しているが,新たな作品に取りかかるごとに様々な問題が発生してくる.特に漢文の扱いに関しては重要な問題であり,何らかの対応策を講ずる必要がある.今後,XML が発展する中で解決することを期待したい.

このようにして作成された日本古典文学本文データベースの XML データが,今後,どのように二次利用,三次利用されていくのか,現段階では予測できないが,さまざまな形で活用されていくことを願ってやまない.

(1) 安永尚志:「国文学研究とコンピュータ」勉誠社 1998.2,pp456-489
(2) 中村康夫: 講座人文科学研究のための情報処理 第 3 巻 テキスト処理編,pp 67-89
安永尚志・丸山勝巳・原正一郎:尚学社, 東京,1998
(3) 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.