4 非専門家指向のデジタル・アーカイヴズに向けて —漢文表現へのXMLの適用—

弘前大学　**内海 淳**

抄録

漢文のデジタル・アーカイヴズはこれまで漢文の出力形式に関して限られた選択肢しか持っていなかったため，非専門家を排除する形になっていた。本稿では，XML（eXtensible Markup Languge）の技術を漢文に適用し，XML技術に基づいたシステムを使用することにより，単一の漢文ソースから，５段階の読みやすさのレベルに対応した，５つの出力形式へと変換できることを示す。このシステムを用いることにより，様々な非専門家の利用者が容易に漢文を利用することができるデジタル・アーカイヴズを構築することが可能になる。

Key Words　デジタル・アーカイヴズ，漢文，XML

Toward Non-Expert Oriented Digital Archives : An Application of XML to Kanbun Expressions

Jun Utsumi

Abstract

Digital archives of kanbun expressions have had very limited options for their outputs, and thus have excluded non-expert users. In this paper, we present an application of XML (eXtensible Markup Languge) to kanbun expressions. We show that, using the XML-based system, we can transform a single source of kanbun into five output formats, which correspond to five levels of accessibility. This system makes it possible to construct a digital archive, in which various non-expert users easily access kanbun expressions.

Keywords: Digital Archives, Kanbun, XML

1　はじめに

歴史資料や文学資料などのデジタル・アーカイヴズ化は，研究者や研究機関などが中心になって行われることが圧倒的に多い。その際にアーカイヴズの利用者として期待されているのは，同じ分野の研究者である。同じ分野の研究者であれば，その興味・関心や資料を利用する能力などは均質であると考えてよいため，その均質さを前提としてアーカイヴズを構築してしまう。このようなデジタル・アーカイヴズは「仲間内のもの」であり，他の分野の研究者や一般の人々が利用できず，閉じてしまっている。誰でも利用できる「開かれた」デジタル・アーカイヴズにするためには，人々の様々な興味・関心や知的レベルに柔軟に対応できるものでなければならない。ここでは，そのような開かれたデジタル・アーカイヴズの可能性を「漢文」を通して考察する。

2　漢文の特性

漢文資料は，日本の歴史や文学において重要な位置を占めている。これまでに多くの研究者や研究機関などが漢文資料のデータベース化や電子化に取り組んできている。しかし，そのような取り組みの中であまり重要視されてこなかった点が２つある。１つは，「日本語としての漢文」であり，もう１つは，「非専門家の利用」である。この２つの問題は密接に結びついている。

ほとんどの日本人にとって，「漢文」は日本語である。我々は，「漢文」の漢字の並びを，一定の規則に従って，意識的に日本語に変換して読んでいる。その日本語への変換を容易にするために，変換の規則を明示的な形で漢字の脇に示したものが訓点である。また，訓点を付しても，難しいと感じる人には，読み下し文という完全な日本語の形が存在する。したがって，多くの日本人にとって，「漢文」は，『日本書紀』や『吾妻鏡』などのような日本国内で書かれたものだけでなく，たとえそれが『史記』などのように中国で書かれたものであっても，日本語の文章として読まれ，受容されているのである。

これまでの漢文資料の電子化，データベース化は，歴史学，漢文学，仏教学などの分野で，主に研究者中心の視点で推進されてきている。これらの分野で漢文を利用する研究者は，高度な漢文読解能力を身につけているため，詳細な訓点や読み下し文などをあまり必要としていない。そのため，これまでの漢文資料の電子化では，原典の校異などの記述に重点が置かれており，非専門家にとっての読みやすさを考慮に入れることは少なかった。

『日本書紀』や『吾妻鏡』などに代表される日本の漢文資料は，日本史や日本文学の研究者だけでなく，一般の人々にとっても関心の高いものであり，デジタル化された場合の利用も多いと予想される。同様に，『史記』

Contact to : utsumi@cc.hirosaki-u.ac.jp

や『漢書』『三国志』のような中国の古典なども，日本語の「漢文」としてアーカイヴズ化が期待されている。これらの漢文を，日本語の資料として提供すると同時に，非専門家の様々なレベルの漢文読解能力に対応した形で提供することが，これからの漢文のデジタル・アーカイヴズ化において重要なことである。

本稿では，このような，漢文の専門家でない人々を視野に入れた，日本語としての漢文のデジタル・アーカイヴズ化について，XML (eXtensible Markup Language) の技術を利用したシステムについて考察する。

3 訓点の扱い

現在，インターネット上で見ることができる漢文の多くが，以下に示すような，白文または読み下し文の形式のいずれか1つ，あるいはこれら2つの形式を併用したものである[1]。

（白文）
送春不用動舟車
唯別残鶯与落花
若使韶光知我意
今宵旅宿在詩家
（読み下し文）
春ヲ送ルニ舟車ヲ動カスコトヲ用ヰず
唯残鶯ト落花とニ別ル
もし韶光ヲシテ我ガ意ヲ知ラ使メバ
今宵ノ旅宿ハ詩ノ家ニ在ラン

漢文に慣れていない一般の利用者にとって，上のような白文を日本語として解読することはかなり困難である。しかし，読み下し文の形で提示されると，ある程度の漢文読解能力を持っていて，原文の白文ないしはそれに近い形を期待している利用者にとっては，不満が残る。このような場合，読み下し文から原文の形を復元することも考えられるが，これにはかなり高度な漢文能力が必要とされる。白文と読み下し文の併用の場合は，その対応関係を1つひとつ確認しなければならず煩雑である。このような事情は，インターネット上に限らず，書籍の場合でもそう変わらない。書籍の場合には，白文ではなく，返り点を付したものが多いが，それでも専門家ではない一般の利用者にとってはかなり難しい。漢文に興味を持った利用者が自分の漢文読解能力にあった資料をなかなか得ることができないでいるのが現状である。

漢文の訓点文，すなわち，白文に返り点などを付したものは，振り仮名や送り仮名の有無に関して，流通している資料ごとにその形態が異なる。しかし，訓点文は，振り仮名や送り仮名の有無によって，日本語としての読みやすさが大きく変わってくる。そこで，ここでは，送り仮名および振り仮名の有無によって訓点文をさらに細分化し，漢文を以下に示す5段階の形式に分けることにする。

(1) 白文：漢字のみから構成されるもの
(2) 訓点文1：漢字＋返り点のみ
(3) 訓点文2：漢字＋返り点＋送り仮名
(4) 訓点文3：漢字＋返り点＋送り仮名＋振り仮名
(5) 読み下し文：漢字を返り点に従って並び替え，送り仮名，振り仮名を補ったもの

1から5に下るに従って日本語として読みやすくなっていく。しかし，1つの漢文資料に対しこれら5つの形式をそれぞれ別個に用意しておくことは，多大の労力を必要とするし，資料の同一性を保つことが難しくなる。本稿では，漢文資料をXMLに従って記述することによって，単一のソースからこれら5つ（またはそれ以上）の形式を自動的に生成することができることを示す。

4 漢文のXML化

上に挙げた漢文をXMLに従って記述すると，Fig. 1のようになる[2]。

<kanji>...</kanji>，<furi>...</furi>，<okuri>...</okuri>の中には，それぞれ，漢字，振り仮名，送り仮名が入り，この3つの要素が，<ku>...</ku>の中に入る。訓点のレ点が関わる部分を<reten>...</reten>で囲み，訓点の一，二，三が関わる部分を，それぞれ，<ichiten>...</ichiten>，<niten>...</niten>，<santen>...</santen>で囲んでいる。訓点を付した漢文は，白文を日本語の規則に従って構造化し，タグを付けたものであると考えることができるので，XMLのタグ付けは簡単かつ規則的に行うことができる。

このXML表現を，PerlやXSLT (eXtensible Stylesheet Language Transformations) などの変換プログラムを使用して適切な出力形式に変換していくのである。その出力の手順は以下のようなものになる。

(1) の白文への変換の場合は，<kanji>...</kanji>で囲まれた部分のみを順次取り出して配置する。
(2) の返り点のみの訓点文1の場合は，白文の場合の手順に加えて，次のような操作を行う。<reten>...</reten>で囲まれた2つの要素の最初の要素の最後にレ点を配置する。それ以外の<ichiten>...</ichiten>などの場合は，そのタグで囲まれた要素の最後にそれぞれに対応する返り点を配置す

```
<reten>
 <ku><kanji>送</kanji><furi>おく</furi><okuri>ルニ</okuri></ku>
 <ku><kanji>春</kanji><furi>はる</furi><okuri>ヲ</okuri></ku>
</reten>
<niten>
 <reten>
  <ku><kanji>不</kanji><furi>ず</furi></ku>
  <reten>
   <ku><kanji>用</kanji><furi>もち</furi><okuri>キ</okuri></ku>
   <ku><kanji>動</kanji><furi>うご</furi><okuri>カスコトヲ</okuri></ku>
  </reten>
 </reten>
</niten>
<ichiten>
 <ku><kanji>舟車</kanji><furi>しうしや</furi><okuri>ヲ</okuri></ku>
</ichiten>

<ku><kanji>唯</kanji><furi>ただ</furi></ku>
<santen>
 <ku><kanji>別</kanji><furi>わか</furi><okuri>ル</okuri></ku>
</santen>
<ku><kanji>残鶯</kanji><furi>ざんあう</furi><okuri>ト</okuri></ku>
<niten>
 <ku><kanji>与</kanji><furi>と</furi><okuri>ニ</okuri></ku>
</niten>
<ichiten>
 <ku><kanji>落花</kanji><furi>らくくわ</furi></ku>
</ichiten>

<ku><kanji>若</kanji><furi>もし</furi></ku>
<santen>
 <ku><kanji>使</kanji><furi>し</furi><okuri>メバ</okuri></ku>
</santen>
<ku><kanji>韶光</kanji><furi>せうくわう</furi><okuri>ヲシテ</okuri></ku>
<niten>
 <ku><kanji>知</kanji><furi>し</furi><okuri>ラ</okuri></ku>
</niten>
<ku><kanji>我</kanji><furi>わ</furi><okuri>ガ</okuri></ku>
<ichiten>
 <ku><kanji>意</kanji><furi>い</furi><okuri>ヲ</okuri></ku>
</ichiten>

<ku><kanji>今宵</kanji><furi>こよひ</furi><okuri>ノ</okuri></ku>
<ku><kanji>旅宿</kanji><furi>りよしゆく</furi><okuri>ハ</okuri></ku>
<niten>
 <ku><kanji>在</kanji><furi>あ</furi><okuri>ラン</okuri></ku>
</niten>
<ku><kanji>詩</kanji><furi>し</furi><okuri>ノ</okuri></ku>
<ichiten>
<ku><kanji>家</kanji><furi>いえ</furi><okuri>ニ</okuri></ku>
</ichiten>
```

Fig. 1　例文のXML化

る。

(3) の返り点＋送り仮名の訓点文2の場合は，取り出す要素が<kanji>...</kanji>と<okuri>...</okuri>であり，<okuri>...</okuri>の部分を送り仮名の出力位置に書き出すことを除けば，(2) の訓点文1の場合と同じである。

(4) の返り点＋送り仮名＋振り仮名の訓点文3の場合も，取り出す要素が<kanji>...</kanji>，<furi>...</furi>，<okuri>...</okuri>の3つであり，<furi>...</furi>および<okuri>...</okuri>の部分を，それぞれ，振り仮名と送り仮名の出力位置に書き出す点以外は (2) の訓点文1と同じである。

(5) の読み下し文の場合は，まず，<reten>...</reten>で囲まれた2つの要素の順番を入れ替える。<niten>...</niten>および<santen>...</santen>の囲まれた要素は一時的に格納しておき，<ichiten>...</ichiten>部分を処理する時点で，<ichiten>...</ichiten><niten>...</niten><santen>...</santen>の順に並び替える。<kanji>...</kanji>，<furi>...</furi>，<okuri>...</okuri>の3つ

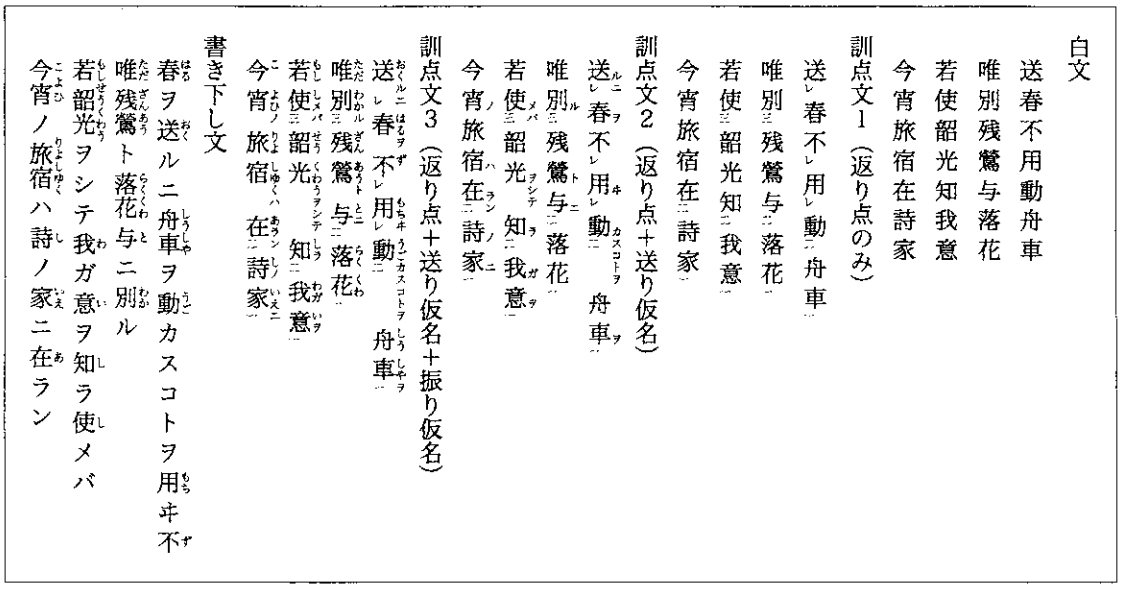

Fig. 2　漢文の出力例

を取り出すが，<okuri>...</okuri>は (3) や (4) の訓点文の場合とは異なった出力位置に配置する。

Fig.1 に示した XML 表現を，この手順に従った Perl のプログラムを通して，出力形式として LaTeX 2ε の形式に変換し，それを組版したものが Fig. 2 である[3]。

この例では，書き下し文の部分に，「用ヰ不」のような，通常は使用されない表現が表れているが，この問題は，書き下し文への変換プログラムをより精緻なものにすることで解決できる。それ以外の点では，我々が書籍などで見る漢文の表現と変わりはない。

しかし，すべての漢文表現が，この XML の形式にすんなり収まるわけではない。例えば，上に挙げた例文の 3 行目に関して，Fig. 3 に示すような，訓点の異なる事例が存在する[4]。

Fig. 3 の訓点文の 3 行目を，上に示した場合と同じ原

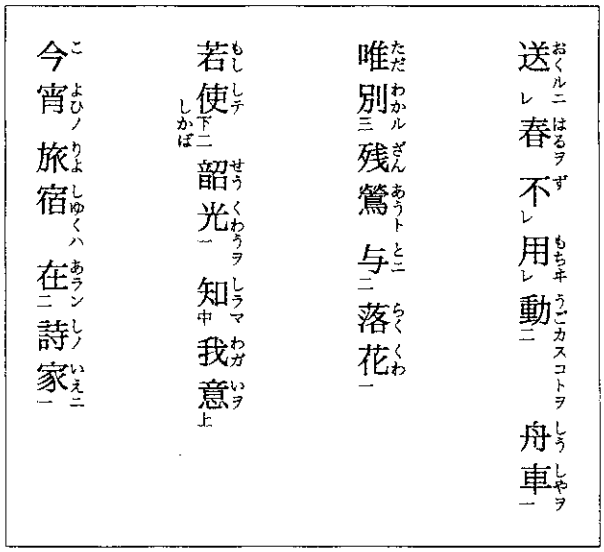

Fig. 3　問題となる訓点の例

```
<ku><kanji>若</kanji><furi>もし</furi></ku>
<niten>
 <sitaten>
  <ku><kanji>使</kanji><furi>し</furi><okuri>テ</okuri><hidari-furi>しかば
</hidari-furi></ku>
 </sitaten>
</niten>
<ichiten>
 <ku><kanji>韶光</kanji><furi>せうくわう</furi><okuri>ヲ</okuri></ku>
</ichiten>
<nakaten>
 <ku><kanji>知</kanji><furi>し</furi><okuri>ラマ</okuri></ku>
</nakaten>
<ku><kanji>我</kanji><furi>わ</furi><okuri>ガ</okuri></ku>
<ueten>
 <ku><kanji>意</kanji><furi>い</furi><okuri>ヲ</okuri></ku>
</ueten>
```

Fig. 4　問題となる訓点文の XML 化

則で，記述するとFig. 4のようになると考えられる。

＜ichiten＞...＜/ichiten＞や＜niten＞...＜/niten＞の部分は，本来，＜ueten＞...＜/ueten＞，＜nakaten＞...＜/nakaten＞，＜sitaten＞...＜/sitaten＞の要素より下位の要素である。しかし，上の表現では，上位の要素である＜sitaten＞...＜/sitaten＞の要素が，下位の要素である＜ichiten＞...＜/ichiten＞の中に生じている。

このような問題を解決するための方策として，まず，訓点の振り方を見直し，XMLの形式に合致するように変更することが考えられる。実際，ここで取り上げた事例に関して，他の資料では，最初に示した例のように，XMLの形式に合致する訓点が振られている。そうした訓点の変更ができない場合でも，XMLの形式や変換プログラムが多少複雑にはなるが，ここで提示している漢文の多様な出力形式への自動的な変換を妨げるものではない。

5 出力形式の問題

次に，これら変換したものをどのような形式で具体的に出力し，利用者に提示するかという問題を考えてみよう。

漢文の出力形式を扱う際に必ず考慮しなければならない事項の1つに，縦書きの問題がある。現在，我々が書籍上で見る漢文はほとんど縦書きである。漢文の訓点文を横書きで提示された場合でも，与えられる情報に違いはないはずであるが，書籍の縦書きの形式に慣れている大半の利用者は横書きの訓点文には違和感を覚え，読みにくく感じるであろう。したがって，縦書きの形式に出力できることが重要になってくる。

もう1つ考慮に入れなければならない事項として，ルビの問題がある。書籍上の訓点文の返り点，振り仮名，送り仮名はルビによって表現されている。したがって，我々が書籍で慣れ親しんでいる形に出力するためには，ルビの機能をサポートしていることが不可欠である。

縦書きの表現，および，ルビの機能の2つを実現している出力形式の1つとして，XHTML(eXtensible HyperText Markup Language)を挙げることができる。現在，インターネット上で主流のHTMLの形式では，これら2つのいずれもサポートされていないため，訓点文の形式を表現できない。HTMLにXMLの拡張性を取り入れて作り上げられたXHTMLは，問題の2つの機能をサポートしているだけでなく，XMLとの親和性の高さという点からも，上に示したようなXMLから様々な漢文の訓点文などを導き出すシステムの出力形式として適している。

縦書きの表現，および，ルビの機能の2つを実現しているもう1つの出力システムとして，LaTeX 2 eを挙げることができる。LaTeX 2 eは，印刷の組版を行うソフトウェアであるが，日本語化されたpLaTeX 2 eでは縦書きをサポートしている。また，LaTeX 2 eでは，単にルビを扱うことができるだけでなく，すでに，藤田眞作氏によって，漢文組版用の優れたマクロプログラム[5]がいくつも作成されており，それらを利用することができる。しかし，LaTeX 2 eの出力システムは，XHTMLの場合より複雑になる。まず，XML形式のソースを変換プログラムを用いて，LaTeX 2 eのソースファイルに変換し，さらにLaTeX 2 eなどを使って実際の出力ファイルを生成する。本来，LaTeX 2 eは紙への印刷が最終出力であったが，現在では，電子文書としての扱いやすさから，PDF (Portable Document Format) 形式で出力することが一般的になっている。

XHTMLとLaTeX 2 eの処理の流れを図式化したものがFig. 5である。

上で述べたように，XHTMLは，XMLとの親和性が高いことから，変換プログラムを介するだけでよく，アーカイヴズ・システムの構成がシンプルになるという利点がある。ただし，現行のHTMLと同様に，紙に印刷する場合には，多少問題がある。また，現時点では，XHTMLに完全に対応しているブラウザがInternet Explorer 5.5以降のものに限られているという問題もあるが，これに関しては，今後，他のブラウザもXHTMLに対応していくと考えられるので，将来的に問題はないであろう。

LaTeX 2 eを介在させる場合は，XHTMLの場合より，多少複雑な構成にならざるをえない。しかし，もともと紙への出力が前提となっているシステムだけに，印刷し

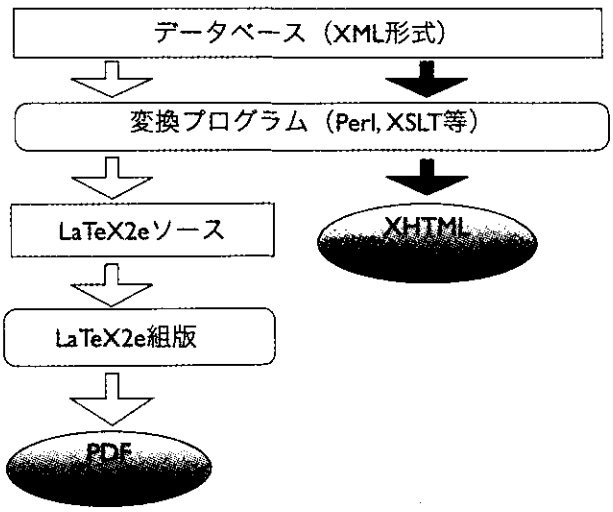

Fig. 5 XHTMLとLaTeX 2 eの処理の流れ

た場合の見やすさはXHTMLより上である。また，PDFはどのOSでも簡単に見ることができる形式であるので，現時点では，この点でもXHTMLより優位である。

いずれの方式を選択した場合でも，利用者はブラウザなどを通して，白文，訓点文，読み下し文のどれを選ぶか，振り仮名や送り仮名は必要かなどをアーカイヴズ・システムに送り，自分の希望する出力形式を受け取ることが可能である。

6 最後に

以上のように，ここでは，利用者の様々な漢文読解能力や関心に応じて，単一のソースから様々な漢文の出力形式を提供できることを示した。漢文のデジタル・アーカイヴズ化は，文字コードという大きな問題を抱えている。しかし，今「漢字ブーム」といわれるほど，漢字に高い関心が寄せられている。このような中，ここに示したような形で，多くの漢文資料が幅広い層の人々に利用できるようになれば，漢文への学習意欲を高め，漢文の普及につながるのではないかと考える。

注

[1] 漢文の例は山下宏明校注『新潮日本古典集成 太平記二』新潮社（1980年）による。

[2] 訓点は，山下宏明校注『新潮日本古典集成 太平記二』新潮社（1980年）によるが，振り仮名を補った。
この表現自体は，完全なXML文書ではない。説明上の便宜のため，文書宣言など，漢文に関連しない部分を省略して提示している。

[3] Fig. 2の組版は，藤田眞作『続LaTeX 2e階梯・縦組編』アジソンウェスレイ（1998年），および，藤田眞作『pLaTeX 2e入門・縦横文書術』ピアソンエデュケーション（2000年）所載のsfkanbun.sty他のスタイルファイルを使用し，pLaTeX 2eで行った。

[4] 3行目の訓点の事例は，岡見正雄校注『太平記（二）』角川文庫（1982年）による。

[5] 注3を参照。

著者略歴
内海 淳（うつみ じゅん）
◎現在の所属：弘前大学人文学部
◎専門分野：言語学，情報教育