デジタルアーカイブにおける異体字を含むテキストの取り扱い方式の提案

来住　伸子†

津田塾大学情報数理科学科

0. はじめに

デジタルアーカイブの重要な目的の一つは、より多くの利用者が資料を利用するようになることである。そのためには、電子データを収集するだけでなく、利用者の興味や知識、利用者の計算機環境に合わせて、データを検索、選択、表示するツールやシステムを整備することが必要である。

1．システム設計の背景と方針

津田塾大学梅子資料室では明治、大正、昭和初期における英語教育と英学に関する資料の電子化を進めている[1]。すでに、数種の資料を画像データとして電子化したが、必ずしも利用者は増えていない。そこで、資料を画像だけでなくテキストとしても電子化し、学生から研究者まで幅広い利用者それぞれに適した表示や検索が可能なシステムの作成を試みることにした。

今回、電子テキスト化の対象としたのは、『英学新報』(英文名　The English Student)という1901 年から 1908 年に発行された、英語を学ぶ学生と英語を教える教師のための雑誌である。この雑誌には当時の英語教育で使用した教材に加え、英米文化社会の紹介や、日本の古典や日本人の著作物の英訳など、さまざまな分野の記事が掲載されている。各分野の記事すべてに興味を持つ利用者は限られるが、個々の記事ごとには、次のような利用者を期待できる。

- 小学生：童話、日本語訳された外国の童話
- 中学・高校生：やさしい英会話、英作文
- 大学生：当時の著名人の投稿や時事解説
- 研究者：日本近代史、言語学、女性学に関連する記事や広告

これらの利用者それぞれに合う形式で資料を検索、表示できるようにするには次のような情報が必要である。

- o 文書の表示構造に関する情報
- o 文書の論理構想に関する情報
- o 使用言語に関する情報
- o 文字に関する情報
- o メタ情報

そこで、これらの情報をＸＭＬタグとしてテキストに付加し、ＸＭＬツールを利用して、検索や表示を行うことを試みることにした。上記の情報を表現するＸＭＬ形式として、すでにいくつかの提案がされている。設計方針として、できるだけ既存のタグを使用することにし、文書の表示構造に関するタグセットは XHTML[2]、メタ情報に関するタグセットとしては OAI[3]を使用した。一方、文書の論理構造は、雑誌記事、雑誌広告、教科書など資料によってかなり異なるため、資料の種類ごとにタグセットを新しく定義した。また、文字情報と言語情報についても独自のタグを使用することにした。既存のタグや属性では、目的とする多様な表示や検索が十分に実現できないためである。また、互換性が必要な場合、独自のタグを変換して、既存のタグセットに合わせることは、ＸＭＬの各種ツールを使用すれば容易であると考える。この報告では、各種資料のうち、雑誌記事に使用するタグを例として紹介する。

2．使用言語に関する情報

使用言語に関する情報とは、ある要素がどの言語で書いてあるかを示す情報である。ＸＭＬでは、ある要素のデータがどの言語で書かれたものかを xml:lang 属性を使用して示すことになっており、その値として使用する文字列も定められている。そのため、どの要素にも xml:lang 属性が使用できるが、日本語であることを指定するには、ja または ja-JP　という文字列しか使用できない。このシステムでは、日本語と英語というような異なる言語で書かれた要素の対応関係を保ちたい、将来は、「明治期の日本語」というように日本語

XML tag set for old Japanese text for digital archives

† Nobuko Kishi, Dept. of Mathematics and Computer Science, Tsuda College

の使用時期ごとに異なる値を使いたい、などの理由から、独自の要素名と属性名を使うことにした。

図１は英学新報中の記事の、画像データの一例を示し、Listing 1 は対応するＸＭＬデータの一部を示す。最上位要素は article 要素で、そこに含まれる最初の要素 metadata にはメタ情報が含まれている。つづく header 要素から画像データに対応する情報が表現されている。

異なる言語で書かれた同じ内容（例：Conversation と会話）を一つの要素 multi とし、その下位の要素（例：text 要素）に属性 lang をつけることにより、言語の種類と対応関係を示している。このページでは、英語が左側、対応する英語が右側に書かれている。このような場合、日本語と英語の対応付けの方法は数通り考えられるが、この例では、text 要素と section 要素の２組が対応付けられている。より下位の要素や上位の要素単位で対応づけることも可能である。

３．文字に関する情報

明治期の文章には常用漢字に含まれない漢字が多く使われているため、現在の利用者が読めない、または読むことが難しい。そこで、元の文章に使われている漢字と、その漢字に対応する常用漢字、または読みなどの代わりの文字列を char 要素として付加することにした。Char 要素は、常用漢字や代わりの文字列をデータとして持ち、原文で使われることの多い、通常のパソコン環境では表示できない可能性のある漢字については、文字セット名と文字コード値で表現することにした。たとえば、「會話」を常用漢字で「会話」と表示したい場合、次のように指定する。

```
<char set="iso-2022-jp" code="98f0">会</char>話
```

これは、原文は、JIS コードで 98f0 に対応する文字を使用し、対応する常用漢字として「会」を使用するということを示している。

明治期の漢字を常用漢字に対応させるには、コード変換を行えばよいという考え方があるが、そうではなく、この char 要素というＸＭＬタグを使用することにしたのはいくつか理由がある。まず、文脈によって、変換してよい文字と変換すべきでない文字がある。たとえば、「女子英学塾は、創立時には女子英學塾と記した」というような文の場合、「學」は常にこのまま、原文で使用された文字で表示すべきである。そのような場合、<char set="iso-2022-jp" code='8feb"/>と指定することにより、変換しない文字であることを示すことができる。また、利用者の使用できる文字セットの大きさによって、使用できる代わりの文字や文字列が異なることがある。たとえば、JIS コード文字セットしか使用できない場合と、Unicode 文字セットも使用できる場合では、異なる。

代わりの文字を XML タグデ指定する方法はすでにいくつか提案されている。JEPAX[4] に使われている gi 要素，XKP[5]で使われている gaiji 要素である。gi 要素は代わりの文字をデータではなく属性値として指定する、gaiji 要素は代わりの代わりの字という指定ができない、などの点が char 要素と異なる。ＨＴＭＬへの変換の容易さを考慮して、独自形式の char 要素を採用することにしたが、実際には、これらの要素は、ほとんど同じ情報を持っているので、形式の変換はＸＭＬタグ変換ツールで容易にできると考える。

４．使用する文字コードセット

明治期の文章には JIS 第１水準や第２水準に含まれない文字が多く使われている。これらの文字の入力と表示には、今昔文字鏡[6]を使用することにした。入力ツールが低価格、Unicode 文字セットを含む、文字フォントを持たない利用者が画像データとして Web ブラウザに文字を表示させることができる、などの理由による。たとえば、「新燿」を常用漢字で「新調」と表示したい場合、次のように入力する。

```
新<char code="055086" set="mojikyo">調</char>
```

「燿」には Unicode 文字セットによく似た字体「瀧」がある。それらを代わりに使用してもよい場合は、次のような表現を使う。

```
新 <char code="055086" set="mojikyo"><char
code="035069" set="mojikyo">調</char></char>
```

５．表示システムと検索システム

上記の方針で設計した XML タグセットを使用し、英学新報のテキスト入力作業を進めると同時に、表示と検索システムの試作を行っている。

表示システムは、XML パーサー xerces [7]を利用して、XML から HTML への変換を行うことにより、Web ブラウザでテキストを読めようにしている。図２から図４は、その表示例である。図２は常用漢字だけで表示した画面、図３はＪＩＳ第２水準までに含まれる異体字で表示した画面、図４は今昔文字鏡を利用して原文にできるだけ忠実に表示した画面を示す。

検索システムは、常用漢字でテキスト表現を統一した後、全文検索システム namazu[8] と XML データベース Xindice[9] を使用して検索を行う。図５は検索システムの入力画面例である。

６．まとめと今後の課題

明治期の文書を全文テキスト入力するためのＸ

MLタグセットを提案した。このタグセットを利用すると表示システムや検索システムが既存のオープンソースソフトウェアを利用して、容易に作成できることを確認した。

今後、実際に利用者に利用してもらうことにより、評価を行う必要がある。次のような課題がまだ残されていると予想している。

- 異体字に関する情報の整備
- 地名・人名の表記方法の整備
- 旧かな遣いへの対応
- 語単位の表記の違いへの対応（例：希臘とギリシア）

今後、より多くの資料の電子テキスト化や、他の組織との協力を通じて、これらの課題に取り組みたい。

参考文献

[1] 来住:Web における外字の取り扱い方式の提案,第 42 回 プログラミング・シンポジウム報告集, pp.127 ? 134, 情報処理学会,2001年1月

[2]http://www.w3.org/MarkUp/

[3]http://www.openarchives.org

[4]http://www.jepa.or.jp

[5]http://www.xkp.or.jp

[6]http://www.mojikyo.org

[7]http://xml.apache.org/xerces2-j/

[8]http://www.namazu.org

[9]http://xml.apache.org/xindice/

Listing 1

```
<?xml version="1.0" encoding="Shift_Jis" ?>
<article  xmlns="http://www.tsuda.ac.jp/arcives/dtd/article"
    xmlns:h="http://www.w3.org/1999/xhtml">
<metadata>
 <oai_dc:dc     xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
     xmlns:dc="http://purl.org/dc/elements/1.1/"  .... >
   <dc:title> The English Student </dc:title>
   <dc:description> Volume 1, Number 1, Article 4   </dc:description>
   <dc:date>1901-11-05</dc:date>
   <dc:type>article</dc:type>
   <dc:identifier>pl001004</dc:identifier>
 </oai_dc:dc>
</metadata>
<header>
<multi>
 <text lang="en">CONVERSATION</text>
 <text lang="jp">
 <char set="iso-2022-jp" code="98f0">会</char>話
 </text>
</multi>
</header>
<text>"Conversation is the vent of character as well as of thought."</text>
<author>EMERSON.</author>
<multi>
  <section page="8" PDFpage="10" lang="en">
  <header>THE NEW BICYCLE.</header>
  <text>
  <x:p>Taro,-Good morning, Jiro! I stopped to see whether your new bicycle
had come yet.</x:p>
  <x:p>Jiro,-Yes, it came yesterday. Won't you come to the stable and take a
look at it? It's a beauty, I think.</x:p>
  <x:p>Taro,-What kind of a wheel is it)
    ..........................
  </section>
  <section page="8" PDFpage="10" lang="jp">
  <header>
  新
  <char code="055086" set="mojikyo">調</char>
  自
  <char code="e77a" set="iso-2022-jp">転</char>
  車
  </header>
  <text>
  <x:p>
  太
  <char code="039431" set="mojikyo">郎</char>
  次
  <char code="039431" set="mojikyo">郎</char>
  さん、お早う、モウ君の自
  <char code="e77a" set="iso-2022-jp">転</char>
  車が
  <char code="98d2" set="iso-2022-jp">来</char>
  たかドウだか、一寸寄つて見たよ。
  </x:p>
  ..........
  </section>
</multi>
..........
</article>
```

CONVERSATION.

THE NEW BICYCLE.

Taro,—Good morning, Jiro! I stopped to see whether your new bicycle had come yet.

Jiro,—Yes, it came yesterday. Won't you come to the stable and take a look at it? It's a beauty, I think.

Taro,—What kind of a wheel is it?

Jiro,—It's a Rambler. Father thought that was as good a wheel as any he could find in Tokyo.

Taro,—Have you tried it yet?

Jiro,—Of course I have. I have ridden all around the yard.

Taro,—Why don't you try it on the street?

"Conversation is the vent of character as well as of thought."

EMERSON.

会　話

新調自転車

図１．画像データ例

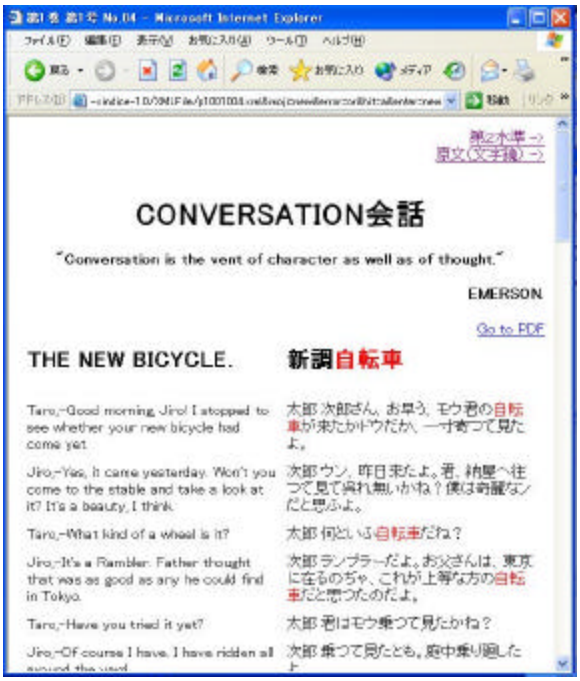

図２．常用漢字を使ったテキスト表示

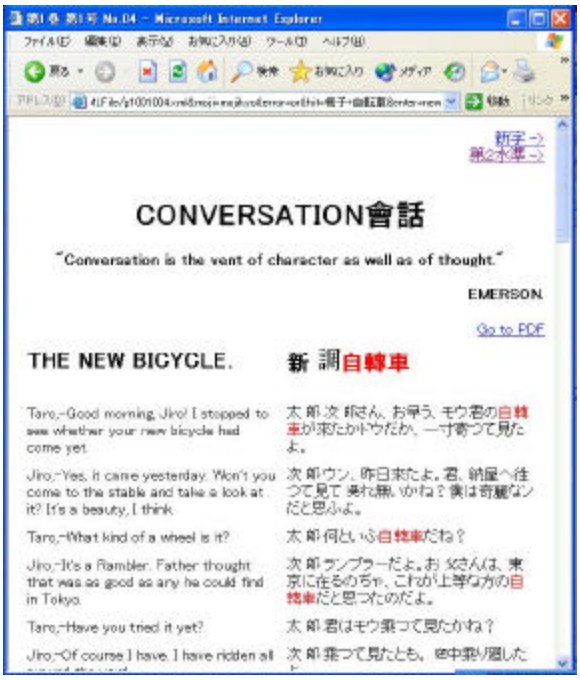

図４．今昔文字鏡に含まれる異体字を使ったテキスト表示

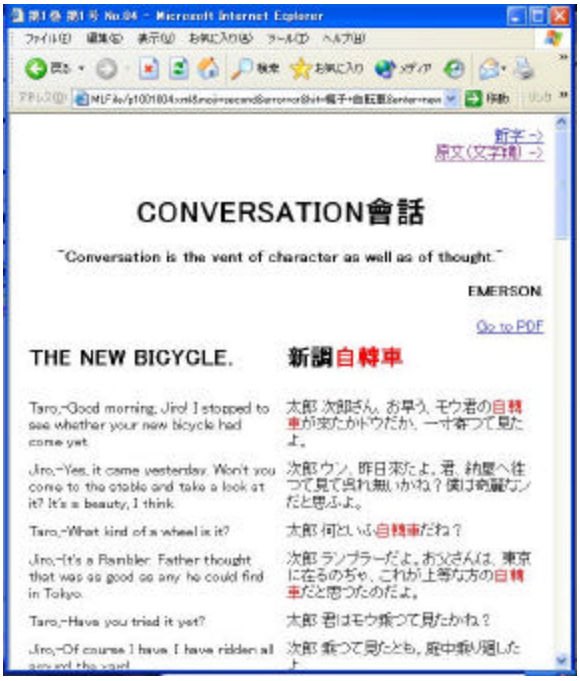

図３．JIS 第２水準にある異体字を使ったテキスト表示

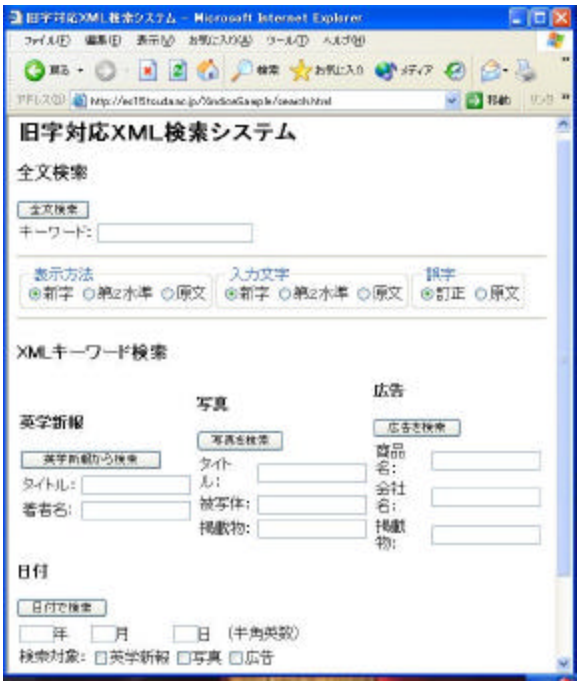

図５．検索システムの入力画面例