連載　第1回　TEIガイドラインの解説

人文学におけるテキストデータ研究活用のための国際標準：TEIのご紹介

一般財団法人人文情報学研究所　永崎 研宣（ながさき きよのり）

はじめに

本誌は、文書情報マネジメント一般に関する雑誌であり、読者の方々は幅広いご関心をお持ちだろう。しかしながら、本稿では、人文学のための文書マネジメントの一環とも言える、テキストデータの構造化という、ややニッチな話をさせていただきたい。ただ、ニッチとはいえ、技術としては大きな広がりと応用可能性を持つ話であり、また、これから本稿で紹介していくような形で人文学のためのテキストデータが作成・公開されていくことで、文書マネジメントの在り方にも良い影響を与えていく可能性もあるかもしれない。そのような観点から、本稿をご高覧いただけるとありがたい。

というわけで、本題に入ろう。文学や歴史、哲学をはじめ、さまざまな分野を含む人文学において、テキストデータと言えば、まず、研究資料をテキストデータ化することで活用する際に有用性が高いと考えられることが多い。また、論文を書く段階になれば、MSワードや一太郎、あるいはLaTeX等を使ってテキストデータを作成する人が多いようである。そのようなことで、テキストデータの作成をまったくしたことがないという人文学研究者はそろそろ極めて希少な存在となっている頃だろう。

テキストデータの作成は、最初から大きな野望を持って取り組む人もいるかもしれないが、とりあえず手元にデータを作っておいて、論文や各種原稿を書くときにコピペして使ったり、ちょっと検索してみたりするのに便利だからと作成する人も多いと思われる。そうこうしているうちに、これがたまってくると、大規模テキストデータを検索することの利便性に気がついたり、そこで他の人が作ったデータとも連携できるようにしたくなったりすることもあるだろう。そして、そのようにしてまとまった大きなテキストデータから、気になる箇所を適宜取り出して索引や表を作ってみたり、人名や地名だけを取り出してみて関係や距離をプロットしてみたり、和歌等の韻律詩であれば韻律を踏まえたデータの分析をしてみたくなったり、文章に登場する年代だけを取り出して時系列に並べ直したりしたくなることもあるかもしれない。

そのようにしてテキストデータを大規模に取り出して便利な使い方をしようと思った場合、近年ではAIに頼るのも徐々に現実的になってきている。実際のところ、大まかな話であればAIでもかなりのことができるようになってきた。しかし、細かく専門的な事柄になると、精度の面ではまだ改善の余地が大きい。人手では扱えないような圧倒的に膨大な量であれば、少々精度に問題があってもAIに頼ってしまうという方向が今後は出てくるかもしれないが、人力で、すなわち、人がテキストを作成しながら注記をしていき、それらを集約する形で色々なデータをうまく取り出したり処理したりできるなら、精度に関する心配がやや薄れるかもしれず、また、その責任の所在が明確化できるという意味でも一定の有用性があると思われる。

テキストデータに注記する手法

ということで、人力でテキストデータを作り注記をつけていくという話題に入ろう。注記を付けるにあたっては、何らかの記号を使ったりタグをつけたり、色々な方法がある。LaTeXのタグは広く用いられているし、最近はMarkdownも広く使われるようになっている。書式を整えるような事柄であれば、そういったものでも十分なことも多い。また、XMLのタグを利用する方法もさまざまな形で広く受容されている。最近の有名なところでは、マイクロソフトのMS-Office（図1）や電子書籍のためのePubなどでは、ユーザ側からはあまり見えないようになっているものの、データとしてはXMLのタグで書式を記述しており、それだけでもユーザは膨大な数になるだろう。

XMLは著名な国際標準規格の一つである。これは、利用者が自由にタグを設定できる仕様であるため、ユーザグループ、あるいは企業などが自分達にとって便利なタグのセットを設定し、それを共有することで利便性を高めるというのが一般的である。MS-OfficeやePubも、そのようにして一定のグループの中で共

```
<w:body>
    <w:p w14:paraId="1A83BD1D" w14:textId="6F1E86AC" w:rsidR="000B53FA"
        w:rsidRDefault="00852135">
        <w:r>
            <w:rPr>
                <w:rFonts w:hint="eastAsia"/>
            </w:rPr>
            <w:t>人文学におけるテキストデータ研究活用のための国際標準：TEIのご紹介</w:t>
        </w:r>
    </w:p>
```

図1　MS-Office（MS-Word）でのXML利用の一例：この原稿のXMLデータ

有すべく設定されたものであり、さらに国際標準規格として策定され広く用いられるに至っている。そして、より用途を絞り込むことで専門的な利便性を高めるべく、XMLのサブセットは実にさまざまなものが策定され利用されている。そのようなサブセットの中の一つとして、ここで採りあげようとしているTEI（Text Encoding Initiative）ガイドラインが人文学研究者の間で利用されているのである。

TEIの始まり

TEIガイドラインは、XML以前から存在したものであり、むしろ、その知見がXMLの策定時に活かされたという類のものである。すでに1980年代、上述のような理由により、人文学研究者達は自らのテキストデータ構築に共通ルールが必要であることを痛感し、この課題に関心を持つ欧米を中心とした人文学研究者達がいくつかの研究助成団体の支援を受けて1987年にニューヨーク州ポキプシーに集結し、その会合において開始されたものがTEIなのである。当初は、昔からマークアップ言語に取り組んでいる方ならご存じの、SGMLというマークアップ言語に準拠したものとして開始された。

この策定にあたっては、データを共有するために必要な規格とはどういうものかについて、深い議論が交わされたようである。まず、いわゆる厳密な規範的標準規格として制定するのではなく、ガイドラインという位置づけにすることが決められた。というのは、人文学にはさまざまな分野と多様な方法論があり、必要となる注記の付け方、すなわちデータ形式もまた多様なものとなる。さらに、人文学が常に方法論を発展させ続けていくという側面を持っているため、データ形式もまた拡張可能でなければならない。このようなことから、TEIはガイドラインとしてデータ形式に関する提案を行うものという位置づけとなった。それを踏まえた上で、この最初の会議での合意事項は「ポキプシー原則」として公表された。これはとても興味深く、かつ簡潔なものなので、以下に引用したい。

> 1987年11月13日，ニューヨーク，ポキプシー．
>
> 1. ガイドラインは，人文学研究におけるデータ交換のための標準的な形式を提供することを目指す．
> 2. ガイドラインは，同じ形式でテキストのデジタル化をするための原理を提案することも目指す．
> 3. ガイドラインは，以下のことをすべきである．
>   1. 形式に関して推奨される構文を定義する．
>   2. テキストデジタル化のスキーマの記述に関するメタ言語を定義する．
>   3. 散文とメタ言語の双方において新しい形式と既存の代表的なスキーマを表現する．
> 4. ガイドラインは，さまざまなアプリケーションに適したコーディングの規則を提案するべきである．
> 5. ガイドラインには，そのフォーマットにおいて新しいテキストを電子化するための最小限の規則が入っているべきである．
> 6. ガイドラインは，以下の小委員会によって起草され，主要なスポンサー組織の代表による運営委員会によってまとめられる．
>   1. テキスト記述
>   2. テキスト表現
>   3. テキスト解釈と分析
>   4. メタ言語定義と，既存・新規のスキーマの記述．
> 7. 既存の標準規格との互換性は可能な限り維持されるだろう．
> 8. 多くのテキスト・アーカイブズは，原則として，交換形式としてのそれらの機能に関して，そのガイドラインを支持することに賛成した．私たちは，この交換を効率化するための

ツールの開発を援助するよう，支援組織に働きかける.

9. 既存の機械可読なテキストを新しい形式に変換することとは，それらの規則を新しい形式の構文に翻訳するということを意味しており，まだデジタル化されていない情報を追加する必要はない.

このうち、６のガイドラインの制定の仕方については、当初はこれに沿って行われたものの、2000年にTEI協会が設立され、それとともにより民主的な手続きに移行した。それ以外の点については、現在もほぼこの方針に沿ってガイドラインの継続的な改訂が進められているようである。

TEIの初期の話に戻ろう。当時はコンピュータがあまり速くなく、ネットワークでのデータのやりとりも容易ではなかったため、あまり複雑なことはできなかったようだが、それでも欧米言語での古典籍や言語コーパスなどのテキストデータベースを作成する際に利用され、徐々に広まっていったようである。ただ、日本での利用については、文字コードの相違の問題が壁として大きく立ちはだかっており、それに加えてネットワークでのデータ共有もまだそれほど容易ではなかったため、当時はあまり広がりを見せることはなかった。

XMLへの移行

その後、TEIを策定した中心メンバーの一部がXMLの議論にも中心的に関わるという形でXMLが策定されたことで、2002年にはTEIもXMLに準拠するものとして書き換えが行われた。当初はSGMLからXMLへの比較的単純な置き換えとなり、XMLの有用性を十分に発揮できなかったものの、その後2007年には本格的なXML対応版としてTEI P5 Guidelinesが公開された。これ以降は、このTEI P5のマイナーバージョンアップという形で改訂されていくことになる。XMLは、ちょうどWebが広まりつつあるなかでさまざまな活用可能性が期待されたこともあり、IT企業や開発者の間に広く圧倒的な速さで受容された。50代以上の方々であれば、その頃は日本でも書店にXMLの解説書が複数平積みされていたことを覚えておられるかもしれない。結果として、それまで採用していたSGMLに比べると、学習コストが下がり、発注に際しても受注可能な企業・開発者が増え、対応するソフトウェアも多かった。このようになってくると、TEIでのデータ作成手法を教育する場合にも、XMLという汎用性の高い技術を教えることができるため、学習者にとってのメリットも大きくなり、大学等でのカリキュラムにも組込みやすくなった。欧米の大学の人文学においてデジタル技術を教育する際の標準的な内容の一部としてTEIが組込まれたことも自然な流れだったと言えるだろう。そして、これを一つの核として、デジタル・ヒューマニティーズ（人文情報学）と呼ばれる学際的分野が欧米では確立していくことになる。このように、XMLに対応したインパクトは非常に大きく、本格的XML対応版であるP5の登場と、人文学におけるデジタル技術活用の進展があいまって、TEIは欧米の人文学においてはデファクト標準として広く利用されることになるのである。

『校異源氏物語』を用いたTEIの事例

では、TEIを利用した事例としてどのようなものがあるのか、それを少しみてみよう。源氏物語の写本を集成し、脚注を用いて相違箇所を提示した資料として『校異源氏物語』という本がある。これは、校訂テキストや学術編集版と呼ばれることもあるものであり、TEIが得意とする分野の一つである。この脚注の相違箇所（ここでは校異情報と言う）をTEIに準拠したXMLマークアップを行うと以下のようになる（図2）。これをよく見ると、<app>タグで始まり、最後の行には</app>がついている。これは、校異情報であることを表す<app>タグの開始タグと終了タグで囲んだということになり、すなわち、この中に校異情報が記述されることになる。校異情報は、主に、底本と呼ばれる主に依拠することになる資料とは異なる記述がある場合に、それ

```
<app>
<lem wit=”#校異 #青池 #青横 #青肖 #青三 #青大 #河系 #河宮 #河尾 #河爲
 #河平 #河大 #別本 #別御 #別國 #別麥”>つけても</lem>
<rdg wit=”#別陽”>つけつゝもやすからぬ事おほく思ひつむるまゝに</rdg>
</app>
```

図2　『校異源氏物語』の校異情報の例

01 きりつぼ

Bibliographic Information

01 きりつぼ

Original Source

池田亀鑑
校異源氏物語

Witness List

- **Witness 校異:** 校異源氏物語
- **Witness 青池:** 青表紙本系 池田本
- **Witness 青横:** 青表紙本系 横山本
- **Witness 青肖:** 青表紙本系 肖柏本
- **Witness 青三:** 青表紙本系 三條西家本
- **Witness 青大:** 青表紙本系 大島本
- **Witness 河系:** 河内本系
- **Witness 河宮:** 河内本系　高松宮本
- **Witness 河尾:** 河内本系　尾州家本
- **Witness 河爲:** 河内本系　爲家本
- **Witness 河平:** 河内本系　平瀬本
- **Witness 河大:** 河内本系　大島本
- **Witness 別本:** 別本
- **Witness 別御:** 別本 御物本
- **Witness 別國:** 別本 國冬本
- **Witness 別陽:** 別本 陽明本
- **Witness 別麥:** 別本 麥生本

Electronic Edition Information:

Responsibility Statement:

Version 校異: 校異源氏物語

いつれの御時にか女御更衣あまたさふらひ 給ける なかにいとやむことなきゝは
にはあらぬかすくれて時めき給 ありけり はじめ より我はと思あかり 給へる 御 方
く めさましきものに おとしめ そねみ給 おなしほと それより下らうの更衣たち
はまして やすからす あさゆふの 宮つかへ に つけても 人の心を のみ うこかし うら
み をおふつもりにや ありけむ いとあつしくなりゆき もの心 ほそけ にさとかち な
るをいよく あかす あはれなる物に おもほして 人 のそしりをも え はゝからせ給
はす世のためし に もなりぬへき御もてなし也 かんたちめ うへ人 なとも あいなく

Version 河平: 河内本系　平瀬本

いつれの御時にか女御更衣あまたさふらひ 給ける なかにいとやむことなきゝは
にはあらぬかすくれて時めき給 ありけり はじめ より我はと思あかり 給へる 御 方
く めさましきものに おとしめ そねみ給 おなしほと それより下らうの更衣たち
はまして やすからす あさゆふの 宮つかへ に つけても 人の心を のみ うこかし うら
み をおふつもりにや ありけむ いとあつしくなりゆき もの心 ほそけ にさとかち な
るをいよく あかす あはれなる物に おもほして 人 のそしりをも え はゝからせ給
はす世のためし に もなりぬへき御もてなし也 かんたちめ うへ人 なとも あいなく
めをそはめ つゝ いとま は ゆき人の 御 おほえ なりもろこしにもかゝることのおこ
りにこそ 世 も みたれあし かりけれと やうくあめのした にもあちきなう人のも
てなやみくさ になりて楊貴妃のためしもひきいてつへく なりゆく にいとはした
なきこと おほかれと かたしけなき 御心はへのたくひなきをたのみにてましらひ
給 ちゝの 大納言はなくなりて はゝ北の方なん いにしへの 人のよしあるにて おや
うちくしさしあたりて 世のおほえはなやか なる御方くにもいたう おとらすな
にことの きしきをも もてなし たまひけれと とりたてゝはかくしき うしろみし
なけれは 事ある時 は なを より所なく心ほそけ也さきの世にも御ちきりやふかか
りけむ 世になく きよら なるたまのをのこみこさへ うまれ 給ひぬいつしかと 心も
となからせ給て いそきまいらせ て 御覧するにめつらかなるちこの 御かたち なり
一のみこ は 右大臣の女御 の御はらにて よせ をもくうたかひなきまうけの君と 世
に もてかしつきゝこゆれと この御にほひにはならひ給 へくも あらさりけれはお
ほかた の やむことなき おほむ思はかりにて この 君をは わたくし 物に お もほ し かしつき
給 事かきりなし はしめより をしなへての うへみやつかへなとしたまふべき ゝはには あらさ
りきおほえいとやむことなく上すめ かしけれと わりなく まつはさせ 給あまりに
さるへき御あそひの おりく なにことにも ゆへある ことのふしく には まつ ま
うのほらせ 給ある 時には おほとのこもり すくして やかて さふらはせ 給ひ なと あ
なかちにおまへさらすもてなさせ 給し ほとにをのつからかろき 方にも みえしを

Version 別陽: 別本 陽明本

いつれの御時にか女御更衣あまたさふらひ 給ける なかにいとやむことなきゝは
にはあらぬかすくれて時めき給 おはしけり はじめ より我はと思あかり 給へる 御 方
く は めさましきものに おもひをとしめ そねみ給 ける それより下らうの更衣たち
なとは あさゆふの 宮つかへ に つけつゝもやすからぬ事おほく思ひつむるまゝに 人の心を
うこかし なけき をおふつもりにや いとあつしくなりゆき て もの心 ほそけ にさとかち な
るをいよく あかす あはれなる物に おもほして 人 のそしりをも はゝからせ給
はす世のためし と もなりぬへき御もてなし也 かんたちめ うへ人 なとも あいなく

図3　タグ付けされたTEIデータをVersioning MachineでHTMLに変換・表示

を本文と対比できるように記述されるものである。ここでは、底本のテキストは<lem>というタグを付され、異なる記述は<rdg>というタグが付けられている。この二つのタグは、wit="…"という属性が付けられているが、これは、「その記述が行われている資料（この場合は写本）がどれか」を略称で示すものである。

このようなタグの付け方に対応して、これをきれいに表示してくれるソフトウェアが米国メリーランド大学のサイトで公開されている。Versioning Machineというソフトウェアであり、Unicodeが十分に普及した現在では、こういった汎用ソフトウェアを日本語テキスト向けに利用することも問題なく可能になっている。そこで、このソフトウェアで上記の箇所を表示してみたのが図3である。

これはTEIの利活用のほんの一例に過ぎないが、これだけでも、その有用性の一端を垣間見ていただけただろうか。この例からは、以下のようなことが言えるだろう。

- ・著名な国際標準規格XMLを踏まえた人文学のための注記の方法が提供されている。
- ・人文学における特定の個別の分野のための注記の記述方法が提供されている。
- ・注記対象の性質が似通っていれば、国・言語に関わらず同じ形式で注記を記述できる。
- ・言語に関係なく、TEIに準拠して作成されたテキストはTEIに対応したソフトウェアであればどれでも同様に処理できる。
- ・TEIに対応したソフトウェアはフリーソフトとして自由に利用できるものが公開されている。

この事例からは、TEIに準拠したデータを作成することで、日本語テキストであってもある程度のメリットがあることをみていただけたかと思う。とはいえ、TEIは、ここまで見ていただいたように、主に欧米で発展してきたものであり、日本語テキストへの適用は一朝一夕にできるものではなかった。そこで、次回の記事では、このTEIの日本での対応状況についてご紹介したい。

連載　第2回　**TEIガイドラインの解説**

人文学におけるテキストデータ研究活用のための国際標準：日本での展開

一般財団法人人文情報学研究所　永崎（ながさき） 研宣（きよのり）

前回に引き続き、人文学におけるテキストデータ研究活用のための国際標準、TEIガイドライン（Text Encoding Initiative Guidelines）を紹介したい。今回は特に日本での対応状況について取り上げる。

日本語対応にあたってのIT環境の困難さ

TEIガイドラインが主に欧米の人文学研究者達によって開始されたことは前回述べた通りである。一応、1987年のその開始当初には1名の日本人が参加していた。その一方で、人文系研究者によるテキストデータの作成はその頃から着実に進められており、全国的な研究会が組織される例もあった。そのような中で日本での導入に向けた取組みも行われたのだが、しばらくの間は日本ではあまり広まらなかったようである。

その頃に日本での導入について検討した論文や著作などがいくつか残ってはいるが、肯定的なものばかりではなかった。そもそもUnicodeが十分に普及する前の段階では、コンピュータのソフトウェアのレベルで文字コード間の互換性が十分ではなく、異なる文字コードに準拠して作成されたテキストデータを利用するには無視できないほどのコストが必要だった。

それゆえ、TEIガイドラインのような国際的なガイドラインに準拠してテキストデータを作ったとしても、それを活用するためのソフトウェアは各文字コード間での互換性を確保するためのコストが大きくなってしまう状況であった。残念ながら、当時はTEIガイドラインに準拠したデータの作成が物事を効率化してくれるとは言いがたく、むしろ高コストにならざるを得なかった。さらに、当初のTEIガイドラインにおいて記述言語として採用されていたSGML（Standard Generalized Markup Language）というマークアップ言語は処理が複雑である上に利用者・利用業界が限定的であり、入門書などが充実しているというわけでもなければ、技術者の間でそれほど広まっているわけでもなかった。したがって、TEIガイドラインに準拠してデータを作成するためには、そのようなやや習得が難しく応用性も低い言語を敢えて学ぶ必要があった。このこともまた、TEIガイドラインが日本で広まるにあたっての困難の一つだったと想定される。なお、このような困難さは日本だけに限られた課題ではなく、広くさまざまな言語文化圏において生じ得るものであったことにも留意しておきたい。

IT環境の改善

このような状況が変わってきたのは、Unicodeの普及とXML（Extensible Markup Language）の登場、そしてTEIガイドライン自体のXMLへの移行が大きな契機となったと見ることができる。まず、世界中の文字を一つのコード表に収めるUnicodeについてみてみよう。今日ではまったく当たり前のことなのだが、世界中の文字を一つのコード表に収めるUnicodeに対応したオペレーティングシステムの普及により、異なる利用言語間での文字コードの違いを考慮する必要がなくなったため、さまざまな文字コードを対象としたソフトウェアの作成が、それまでに比べて極めて容易になった。

この効果はすぐには現れず、コンピュータ環境が全体としてUnicodeをベースとするのにやや時間を要したものの、関係各位の尽力により、現在では国際対応として作成された汎用性の高いソフトウェアに日本語処理機能が組込まれるところまで来ているものが増えてきており、日本語テキストの処理にも容易に対応できるものが多くなってきている。逆に日本語向けに作成されたソフトウェアがほぼそのまま他の言語に対応できるようにもなっており、文字コードに起因する障壁はほぼ取り払われたと言っていいだろう。結果として、例えばTEIガイドラインに準拠したテキスト編集ソフトウェアが多様な言語に対応できるようになり、あるいは、TEIガイドライン準拠テキストのタグを利用した分析用プログラムを作ろうとした際にテキストの対象言語にかかわらず同様に利用できるものを簡単に作成できるようになる

など、人文学のためのテキストデータ構築においても文字コードの相違を意識せずに汎用的なものを目指せるようになっている。

一方、XMLの登場とそれへの移行もまた、大きな転換点をもたらした。XMLの登場については前号の連載をご覧いただきたいが、このことは日本においてTEIガイドラインが広まるにあたっても重要な変化であったことを改めて強調しておきたい。

コミュニティとしての対応

IT環境としては徐々に発展してきたものの、一方で、TEIガイドラインがうまく利用されるためにはある程度の規模のユーザコミュニティが必要である。そして、TEIガイドラインの主要ユーザとは、欧米圏においてそうであるように、日本においても人文学研究者である。

XMLに準拠してテキストデータを作成し、それを活用してさまざまな研究を発展させるという営みは、欧米圏においてはすでに一定程度の広がりを見せており、さらに2005年頃からは人文系の大学院生を含む研究者にその種のことを教えるための仕組みとともにデジタル・ヒューマニティーズという領域として着実に広まっていき、さらに学術政策として政府等からも支援されるようになっていく。とりわけ、オープンサイエンス・オープンアクセス・オープンデータといったオープン化政策や研究データ管理のような文脈から国際的な標準規格等に準拠することを研究費助成団体が強く推奨するようになったことがこれを後押ししたようである。すなわち、人文学の資料や成果のオープン化のために必要な国際標準規格への対応が求められたときに、TEIガイドラインは、まさにうってつけだったのである。そこでTEIガイドラインに準拠したデータを作成し公開・共有するという流れが形成され、TEIガイドラインの重要性はより高まっていくことになるのである。しかしながら、日本においてはそのような形で広まるにはさらなる時間が必要であった。というわけで、日本でのコミュニティ形成に向けた動きのようなものを少しみてみよう。

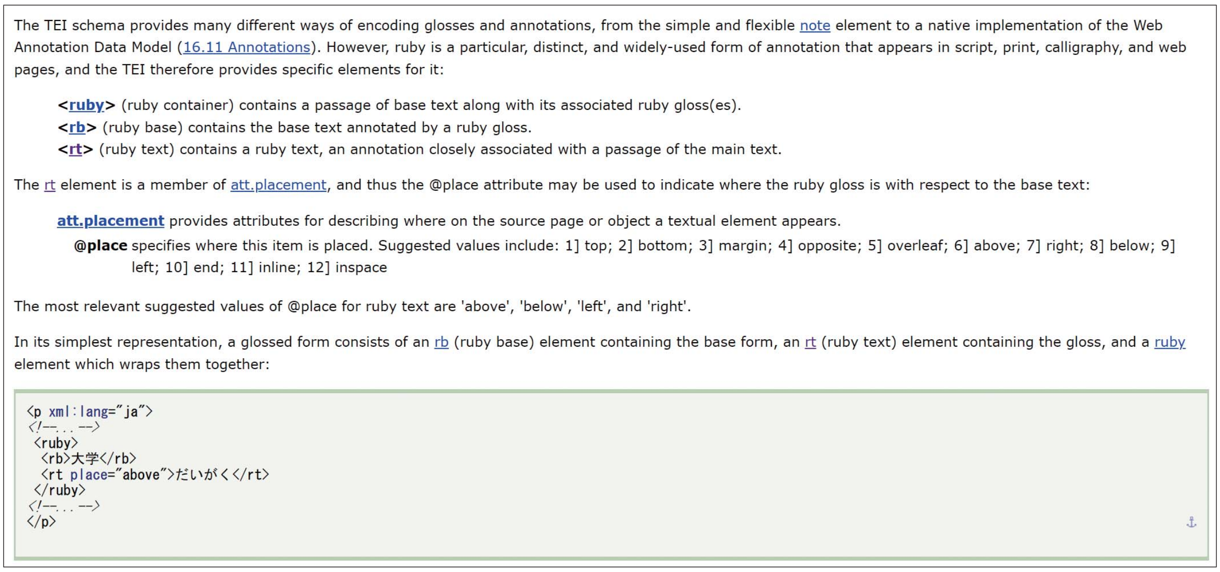
日本におけるTEIに関するコミュニティ形成という観点で注目しておきたいのは、京都大学人文科学研究所において推進された京都大學21世紀COEプロジェクト「東アジア世界の人文情報學研究教育據點」において、Christian Wittern氏が2006年に開催した「国際セミナー TEI Day in Kyoto 2006」である。ここでは、TEIガイドラインについて議論するTEI協会の人々が海外から参集するとともに、日本でごく少数ながら個別に独自のテキストデータ構築やTEIガイドラインに取り組んできた人達が登壇して日本でのTEIガイドライン普及についての検討も行われた。これ自体がすぐに何かにつながったわけではないが、TEIガイドラインを国内の関係者に知らしめる一つの機会にはなったかもしれない。

その後、水面下ではさまざまな動きがあったようだが、現在でも確認できる大きな事柄としては、鶴見大学の大矢一志氏が公表したTEIスキーマの日本語訳がある。これは、TEIタグの入力をする際に日本語でタグの説明を確認することを容易に可能としたものであり、その後の日本での受容に大きな貢献をすることとなった。また、2012年から東京大学大学院人文社会系研究科で人文情報学に関するカリキュラムが開始され、その中でTEIガイドラインを扱う教育も展開されるようになる。そして、これと前後して国内の一部の人文系研究機関が自機関資料の作成や公開にあたりTEIガイドラインの採用を具体的に検討するようになる。こうした流れを受け、TEIガイドラインを管理し改良しているTEI協会において東アジア・日本語分科会が設置されたのが2016年のことであった。

東アジア・日本語分科会の設置

それまでは、TEIガイドラインは本来、抽象的なテキストを扱うものであり、いずれかの特定の言語文化圏に特化した記述手法に対応すべきでないという考え方が強かった。しかし一方で、TEIガイドライン自体は抽象的なものであったとしても、それを具体的に個別のテキストに適用するにあたってはそれぞれの対象資料に即したマークアップ手法が検討されるべきであるという考え方もあった。後者に沿った活動としては、書簡の分科会における書簡に特化したマークアップ手法の検討や、碑文のマークアップに特化したマークアップ手法としてのEpiDoc（Epigraphic Documents in TEI XML）など、いくつかの活発なコミュニティによるものがある。とはいえ、特定の言語文化圏に特化した分科会のようなものは、2016年に至るまで設置されてこなかったようである。

現在のTEI協会には、非ラテン文字（non-Latin）を扱うコミュニティとして、東アジア・日本語分科会とインドテキスト分科会が設置されている。日本のテキスト資料は日本固有のものだけでなく、東アジア文化圏全体の中で作成され発展してきた側面がある。いわゆる漢文が古くからさまざまな場面で用いられてきており、それに返点やヲコト点などのさまざまな記号を付与することで日本語として読めるようにしてきた文化や、さらに、漢字から派生して音を表す仮名文字を用いた文書もまた数多く作成



The TEI schema provides many different ways of encoding glosses and annotations, from the simple and flexible note element to a native implementation of the Web Annotation Data Model (16.11 Annotations). However, ruby is a particular, distinct, and widely-used form of annotation that appears in script, print, calligraphy, and web pages, and the TEI therefore provides specific elements for it:

<ruby> (ruby container) contains a passage of base text along with its associated ruby gloss(es).
<rb> (ruby base) contains the base text annotated by a ruby gloss.
<rt> (ruby text) contains a ruby text, an annotation closely associated with a passage of the main text.

The rt element is a member of att.placement, and thus the @place attribute may be used to indicate where the ruby gloss is with respect to the base text:

att.placement provides attributes for describing where on the source page or object a textual element appears.
@place specifies where this item is placed. Suggested values include: 1] top; 2] bottom; 3] margin; 4] opposite; 5] overleaf; 6] above; 7] right; 8] below; 9] left; 10] end; 11] inline; 12] inspace

The most relevant suggested values of @place for ruby text are 'above', 'below', 'left', and 'right'.

In its simplest representation, a glossed form consists of an rb (ruby base) element containing the base form, an rt (ruby text) element containing the gloss, and a ruby element which wraps them together:

```
<p xml:lang="ja">
<!--. . . -->
 <ruby>
  <rb>大学</rb>
  <rt place="above">だいがく</rt>
 </ruby>
<!--. . . -->
</p>
```

TEIガイドラインにおけるルビの説明の一部

されてきた。そのような過程を踏まえると、日本の歴史的なテキスト資料は東アジアという視点がなければ捉えきれないものであると言え、東アジア・日本語分科会という名称はそのような事情を反映しているとみることもできるだろう。そのような議論を踏まえつつ、Charles Muller（現在は武蔵野大学）氏及び筆者の提案が実り、2016年になってようやく、特定言語に特化された初めての分科会としての東アジア・日本語分科会が設定されたのである。

そして、この論理は他の長い歴史を持つ言語圏にも適用可能であり、特にサンスクリット語に関して活発なユーザコミュニティを有していたインドのテキストに関しても、その後すぐに分科会が設置されることとなった。このことは、日本から始めた議論が他のローカルな言語文化圏における人文学の発展に貢献した事例であり、日本からの国際貢献と言ってよいものだろう。

東アジア・日本語分科会設置後の動向について

東アジア・日本語分科会の設置後は、日本でTEI協会の活動がより公式な形でできることになり、TEIをテーマとするシンポジウムや研究会が開催されるようになっていく。2022年には日本学術会議の公開シンポジウムにおいてもTEIが採りあげられることとなった。

そして、TEIを採用する動きも本格化し、この分科会の設置に尽力したSAT大蔵経テキストデータベース研究会における仏典のTEI化は現代日本語訳の公開を皮切りに着々と進められていく一方で、研究機関としての取組みも、国立歴史民俗博物館における延喜式のTEI化や、国文学研究資料館におけるTEIに準拠した日本古典籍マークアップ手法の開発など、それぞれの立場から取組みが進められてきている。

さらに、この分科会が設置されたことは、欧米中心だったTEI協会において東アジアの言語文化圏の位置付けを問い直すきっかけとなった。

まず2018年には、それまで欧米でしか開催されてこなかったTEI協会の会員総会を東京で開催することとなった。

これは共催した日本デジタル・ヒューマニティーズ学会の参加

TEI研究会は、日本語を含む東アジア言語に関してテキストデータの効果的な構造化の手法について皆で考えつつ学ぶ研究会です。オンラインで開催しておりますので、お気軽にご参加ください。

4/22（土）10:00〜16:00、TEI初心者向け講習会を無料でオンライン開催しました。

近いうちにまた開催しますので、こちらのサイトを時々ご覧ください。

HPから参加申し込みが可能となっている。

者もあわせた数字ではあるものの、300人規模の参加があり、日本での発展を大いに印象づけることとなった。

そして、分科会設置後の長い議論を経て、2020年には分科会の運営委員である岡田一祐氏（現在は慶應義塾大学）、中村覚氏（東京大学）及び筆者が中心となってルビの構造の導入とそれに伴うTEIガイドラインの改訂についての提案書が提出された。そして、その後のさらなる議論の結果、2021年には、日本語のルビの構造がTEIガイドラインに導入されることとなった。なお、この議論の経過等については、TEIガイドラインの改訂に関わる議論自体が基本的にGitHubで行われているため、その記録は現在もGitHubのサイト上で確認できる[※1]。

この分科会の活動は、現在では日本学術振興会の科学研究費補助金の助成を得て毎週開催されており、TEIガイドラインの日本語訳や、さまざまなタイプの資料の試験的なマークアップに取り組んでいる。これまでには、多くの源氏物語写本を集めて相違箇所の情報を本文に即してまとめた池田亀鑑による『校異源氏物語』の校異情報（写本間の相違箇所の情報）やその本文と与謝野晶子訳源氏物語（青空文庫所収）の訳文との文章単位でのリンク、青空文庫で公開されている書簡や戯曲などのマークアップ、方言資料のマークアップなど、さまざまなものが協働で試験的にマークアップされており、そして、その成果の多くはGitHubで公開されている。また、この研究会に関する情報は、研究会Webサイト[※2]があり、今後の予定が告知されているだけでなく、過去の研究会の内容も公開している。参加は誰でも歓迎であり、この種の事柄に興味がおありの方はぜひご参加の申込みをいただきたい。

終わりに

というわけで、今回は、TEIガイドラインの日本での状況についておおまかに見てきた。日本が本格的にTEIガイドラインに対応しようとすることが、結果として、国際的なコミュニティをよりグローバルに有効なものにしていくことに貢献してきたという点が、筆者にとっては大変興味深いものであった。

次回は、TEIガイドラインに準拠して作成されたデータの活用例についてみていきたい。

※1　https://github.com/TEIC/TEI/issues

※2　https://tei.dhii.jp/

連載　第３回（最終回）　TEIガイドラインの解説

人文学におけるテキストデータ研究活用のための国際標準：TEI関連のツールの色々

一般財団法人人文情報学研究所　永崎 研宣（ながさき きよのり）

はじめに

2023年９月４日から８日までの間、ドイツのパーダーボルン大学にて、TEI（Text Encoding Initiative）協会による国際会議が行われた※1。これは基本的には毎年行われるもので、2018年には東京で開催されたこともある。2023年は、TEIから派生した楽譜を符号化するためのガイドラインであるMusic Encoding Initiative（MEI）との共同開催となり、テキストと楽譜の研究者が一堂に会してデジタル化・構造化の議論を行う興味深い場となった。

TEIの国際会議では、テキストデータの構造化やその活用手法についての研究発表が世界中から集結する。本稿のテーマは「ツール」であることからツールに着目してみると、すでにサービスとして提供されているものやベータ版のようなものから単なる試行段階のものまで、さまざまな段階のものが例年のTEI国際会議では発表される。試行段階で発表されたものの結局実を結ばないものもあるが、一方で、十分に成熟し、誰にでも使える段階に達したものは、そのツールやサービスをテーマとするワークショップが開催されることもある。

今回は、テキストを扱うツールとしては、米国メリーランド大学のRaffaele Viglianti氏が中心になって開発されているCETEIceanとGatsby、それから、ドイツのeXist SolutionsのWolfgang Meier氏らによって開発されているTEI Publisherのワークショップが提供されていた。前者はJavaScriptでWebのフロントエンドを担うツールであり、後者はXMLネイティブデータベースであるeXistDBを核にしてTEIテキストをリッチな形で検索・公開するためのミドルウェアである。それぞれに提供された学びの場に多くの参加者達が集っていた。

TEIを変換するための言語

現在のTEIガイドラインはXML（Extensible Markup Language）に準拠した記述方法を採用しているため、記述したデータはXMLのツールチェーンを通して活用することができる。特に、XMLのデータを変換する言語であるXSLT（Extensible Stylesheet Language Transformations）は、その取り組みやすさから、欧米のTEIガイドラインの入門講座においてしばしば教えられるものであり、TEIガイドラインの利用者の間では広く用いられている。XSLTはXMLをHTML等に変換してWebブラウザに表示するだけでなく、簡単な計算も可能であり、後述するように、それを活かしたさまざまなツールが開発されている。

また、多くのプログラミング言語がXMLのデータをその形式に沿って処理できるように専用のライブラリや関数を提供している。近年Webでもよく使われるPHPやJavaScript、Pythonと言ったプログラミング言語はその典型である。PHPの場合は、主にWebサーバ側で用いられることが多い言語であり、サーバ側でTEIファイルを変換してしまう場合に向いている。

JavaScriptは、Webブラウザで動作するプログラミング言語であり、Webブラウザ側で主に用いられるものである。近年はWebサーバ側で用いることも増えてきているが、Webブラウザ上でインタラクティブな機能を提供しようとした場合にはWebブラウザ上でJavaScriptを動作させるべくプログラミングをすることになる。

Pythonは、PHPのようにWebサーバ上で動作させることが可能であり、そのような用途にも使われるが、近年よく使われるのはテキスト分析や統計処理、あるいは深層学習といったものである。Pythonにはそのような機能を実装するためのライブラリが非常に充実しており、比較的容易にそのような処理が可能で

※1　https://teimec2023.uni-paderborn.de/

ある。つまり、手元のパソコン等でテキストを処理・分析したりする際にはPythonが有益ということになる。

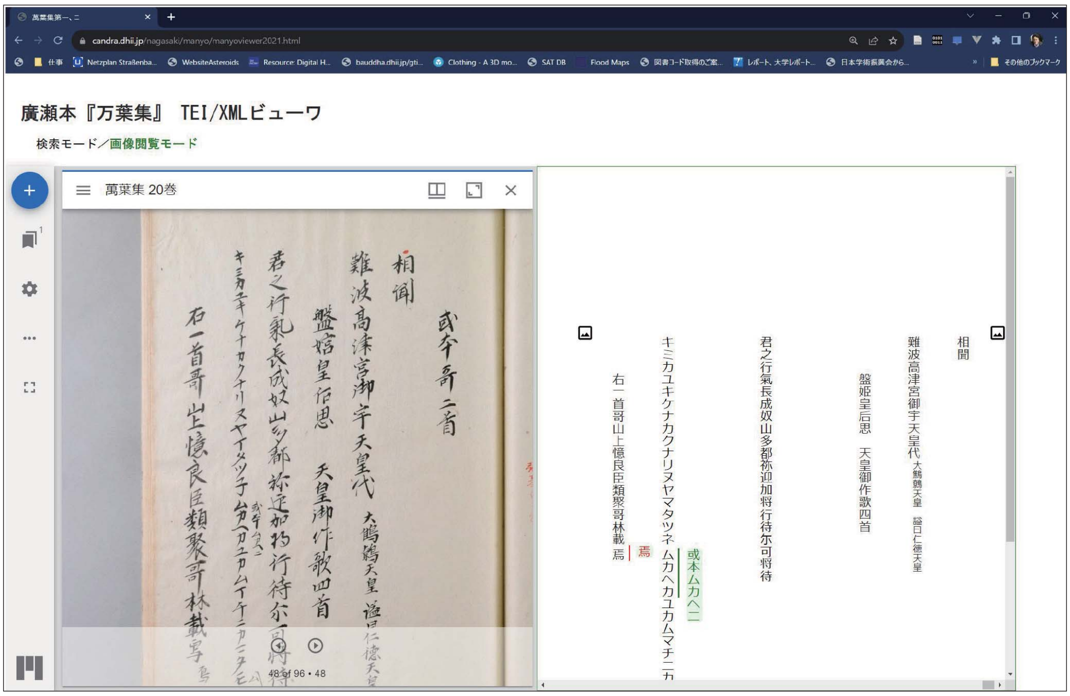
このように、言語ごとにそれぞれ特徴があり、データの状況や処理・公開の仕方に応じて選択することになる。たとえば、小規模のデータを少しだけ公開して便利に使ってもらおうとするのであれば、JavaScriptを用いてWebブラウザ上で表示できるようなものを作成するか、JavaScriptで書かれたツールを利用すれば十分である。しかし、この場合は、処理系はWebブラウザになってしまうため、あまり大量のデータを読み込ませるとWebブラウザがうまく動かなくなったりフリーズしてしまったりすることがある。そこで、一度に処理するデータが大量になる場合には何らかの方法で工夫することになる。たとえば、サーバ側である程度データを分割した上でWebブラウザ側に配信するという方法である。サーバ側でデータを分割するという処理の場合には、PHPやPythonといったプログラミング言語を用いる選択肢も出てくる。さらにデータが大きくなる場合には、TEIのデータをサーバ上のデータベースに格納して取り出すという方法も検討することになる。

TEIを活用するための汎用ツール

上記のようなプログラミング言語を用途に応じて選択しながらTEIの活用ツールは作成されている。汎用的なツールもあれば、個々の資料の構造の特徴を活かした処理ができるよう工夫が凝らされているものもある。

たとえば、上述のCETEIcean[※2]はJavaScriptで書かれたTEIファイルのブラウザ表示用のツールである。これにTEIのファイルを読み込ませればJavaScriptでHTMLに変換してWeb頁として表示できるようになる。それだけでなく、このツールでは、CSSや簡単な関数を追加することでさまざまなカスタマイズが可能となっている。日本語のチュートリアルも公式GitHubで提供されており、入門しやすいものの一つである。

筆者も数年前にこのCETEIceanを改良して『廣瀬本万葉集』の写本のTEIファイルを表示・検索するツールを開発したことがある（図1）[※3]。これらはいずれもデータ量が少ないから利用可能なものだが、さまざまなインタラクティブな機能を追加できる点は特に有用な点である。ただ、この表示の仕方では、Webブラウザ上で変換処理をすることになってしまうため、いわゆる検索エンジンなどにヒットしにくくなってしまうという問題がある。


廣瀬本『万葉集』 TEI/XMLビューワ

検索モード／画像閲覧モード

萬葉集 20巻

相聞

難波高津宮御宇天皇代（大鷦鷯天皇　謚曰仁徳天皇）

盤姫皇后思　天皇御作歌四首

君之行氣長成奴山多都祢迎加将行待尓可将待

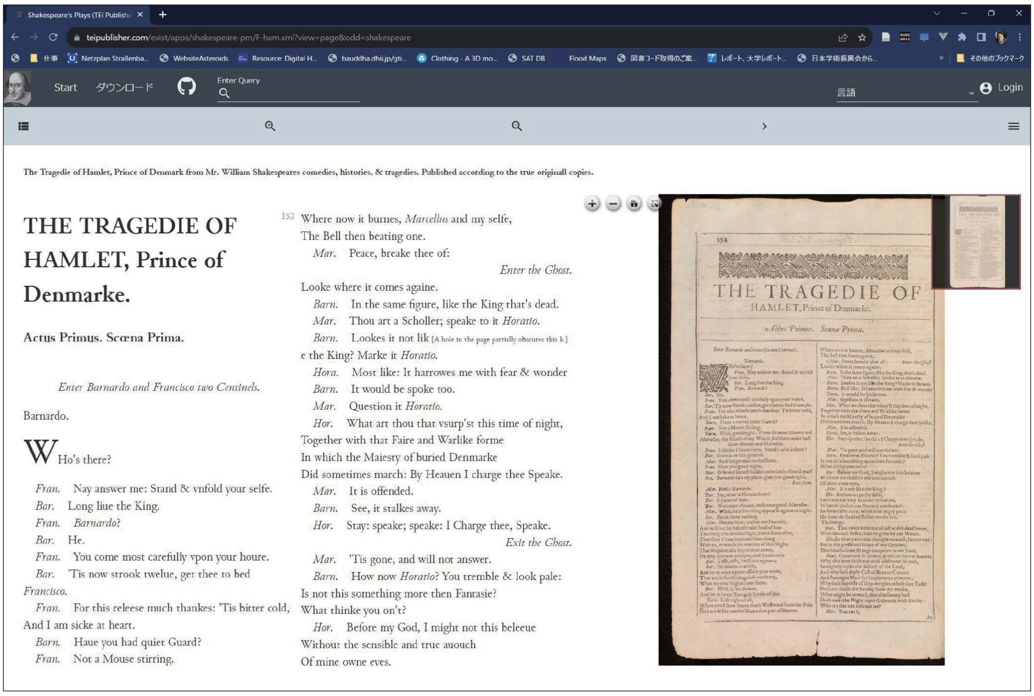
キミカユキケナカクナリヌヤマタツネ（或本ムカヘニ）ムカヘカユカムマチニ

右一首哥山上憶良臣類聚哥林載焉（焉）

図1　CETEIceanの改良版による『廣瀬本万葉集』TEIデータの表示例

[※2] CETEIcean https://github.com/TEIC/CETEIcean/tree/master/tutorial_ja

[※3] 『廣瀬本万葉集』TEI/XMLビューワ https://candra.dhii.jp/nagasaki/manyo/manyoviewer2021.html



The Tragedie of Hamlet, Prince of Denmark from Mr. William Shakespeares comedies, histories, & tragedies. Published according to the true originall copies.

THE TRAGEDIE OF HAMLET, Prince of Denmarke.

Actus Primus. Scœna Prima.

Enter Barnardo and Francisco two Centinels.

Barnardo.

WHo's there?

Fran. Nay answer me: Stand & vnfold your selfe.
Bar. Long liue the King.
Fran. *Barnardo?*
Bar. He.
Fran. You come most carefully vpon your houre.
Bar. 'Tis now strook twelue, get thee to bed *Francisco.*
Fran. For this releese much thankes: 'Tis bitter cold, And I am sicke at heart.
Barn. Haue you had quiet Guard?
Fran. Not a Mouse stirring.

152 Where now it burnes, *Marcellus* and my selfe,
The Bell then beating one.
Mar. Peace, breake thee of:
Enter the Ghost.
Looke where it comes againe.
Barn. In the same figure, like the King that's dead.
Mar. Thou art a Scholler; speake to it *Horatio.*
Barn. Lookes it not lik [A hole in the page partially obscures this k.]
e the King? Marke it *Horatio.*
Hora. Most like: It harrowes me with fear & wonder
Barn. It would be spoke too.
Mar. Question it *Horatio.*
Hor. What art thou that vsurp'st this time of night,
Together with that Faire and Warlike forme
In which the Maiesty of buried Denmarke
Did sometimes march: By Heauen I charge thee Speake.
Mar. It is offended.
Barn. See, it stalkes away.
Hor. Stay: speake; speake: I Charge thee, Speake.
Exit the Ghost.
Mar. 'Tis gone, and will not answer.
Barn. How now *Horatio*? You tremble & look pale:
Is not this something more then Fantasie?
What thinke you on't?
Hor. Before my God, I might not this beleeue
Without the sensible and true auouch
Of mine owne eyes.

図2　TEI Publisherでの『ハムレット』TEIデータの表示例

それを解決するのがGatsbyというプログラムである[※4]。これはMetaとコミュニティによって開発されているユーザインタフェース構築のためのJavaScriptライブラリ、Reactを用いて作成されており、検索エンジンにもヒットしやすい形での公開が可能である。

TEI Publisher（図2）も汎用のツールである。こちらは、XMLネイティブデータベースであるeXistDBにTEIのデータを格納し、XQueryという問い合わせ言語でデータを取り出しつつ表示するところまでを一つのツールでできるようにしているものである。XQuery対応ということでデータの取り出し方の柔軟性が高く、表示・検索のいずれにおいても有用であり、大規模なデータにも対応できる。こちらは導入にある程度の技術力が必要となり、簡単なものとは言えないが、技術的に対応可能であれば有力な選択肢となるだろう。

CETEIceanほどの柔軟性はないものの、JavaScriptを用いて整形・表示を行うツールには、TEI協会東アジア／日本語分科会のGitHubサイトで公開されているものがある[※5]。東京大学史料編纂所の中村覚氏が開発したものが主だが、ここには、発話と固有名詞を抽出して表示するものや、戯曲の台詞をLINE風チャット画面で表示するもの、書簡のデータを地図・年表上に表示するもの、学術編集版テキストの異文情報を読みやすく表示するもの等が提供されている。いずれもあまり大きなデータを読み込ませるとうまくいかないようだが、数メガバイトくらいであればなんとかなるだろう。ぜひ、一通り試してみていただきたい。

学術編集版のためのTEIを活用ツール

こうした汎用的なもの以外にも、個別分野に特化したツールもあり、注目されているものの一つに、学術編集版（やや異なるものだが、校訂テキストと言う場合もある）の異文情報の表示がある。たとえば、写本等で伝承されて複数の少しずつ異なるテキスト資料が現存し、それらの差異を踏まえつつ読解しなければならない文献の場合、異文情報を伴う学術編集版のテキストが作成されることがあり、これはデジタル時代になってデジタル版としてより高機能な形で構築されるようになっている。TEIで広くテーマとなっているものの一つであり、これに対応するツールとしては、有名なものだけでも3つのものが存在する。それぞれに特徴があるので紹介しよう。

一つは米国で開発され公開されているVersioning Machine[※6]である。これは原稿執筆時点ではバージョン5になっており着々と利便性の高い機能を備えつつある。TEIに準拠して作成された学術編集版のテキストにおいて、各対照資料のテキストデータを個別に再構築し対応箇所を示す機能を中心として、注記の表示機能なども充実している（図3）。これはTEIのテキストを変換するためのXSLTのファイルとして頒布されており、Oxygen XML Editorやその他の何らかのXSLT変換ソフトウェア（フリー

※4　Gatsbyを利用する際には以下のURLなどを参照されたい。
https://github.com/raffazizzi/gatsby-ceteicean-starter

※5　https://github.com/TEI-EAJ/aozora_tei/wiki#%E8%A6%96%E8%A6%9A%E5%8C%96%E3%83%84%E3%83%BC%E3%83%AB

※6　http://v-machine.org/

絵入り源氏物語＋校異源氏物語

Bibliographic Information

絵入り源氏物語＋校異源氏物語

Original Source

国文学研究資料館 絵入り源氏 校異源氏物語 青表紙本系 大島本 青表紙本系 肖柏本 青表紙本系 三條西家本 青表紙本系 横山本 青表紙本系 池田本 河内本 河内本系 御物本 別本 別本 國冬本 別本 陽明本 別本 麥生本

Witness List

- **Witness 国絵:** 国文学研究資料館 絵入り源氏
- **Witness 校異源氏:** 校異源氏物語
- **Witness 青大:** 青表紙本系 大島本
- **Witness 青肖:** 青表紙本系 肖柏本
- **Witness 青三:** 青表紙本系 三條西家本
- **Witness 青横:** 青表紙本系 横山本
- **Witness 青池:** 青表紙本系 池田本
- **Witness 河:** 河内本
- **Witness 御:** 河内本系 御物本
- **Witness 別:** 別本
- **Witness 別國:** 別本 國冬本
- **Witness 別陽:** 別本 陽明本
- **Witness 別麥:** 別本 麥生本

Electronic Edition Information:

Publication Details:

Encoding Principles

Version 別陽: 別本 陽明本

n

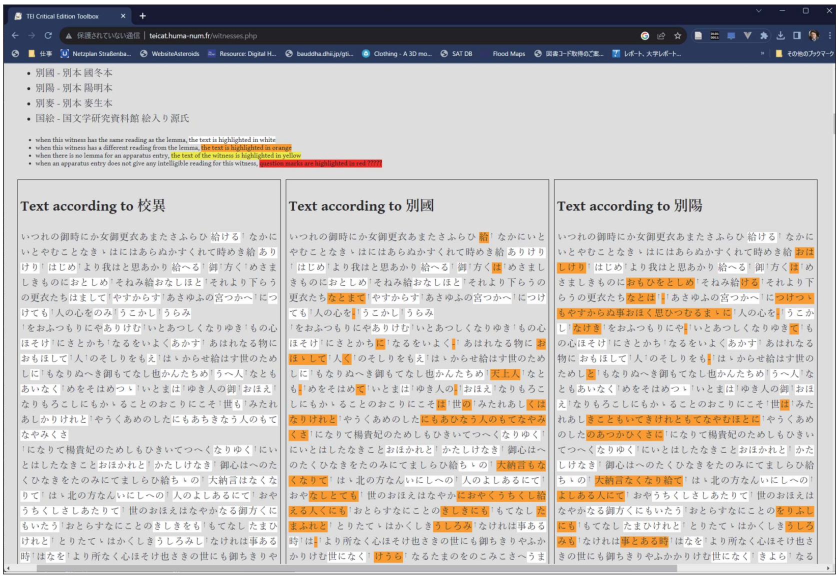
いづれの御時にか女御更衣あまたさふらひ 給ける
なかに。いとやむごとなききはにはあらぬが。
すぐれてときめき給ふ おはしけり はじめ よりわれ
はと思ひあがり給 へる 御 かた／"＼ は めざましきもの
に おもひをとしめ そねみ給 ける 。それより
下らうの更衣たち なとは あさゆふ
のみやづかへに つけつゝもやすからぬ事おほく思ひつむ
るまゝに。人のこゝろを うごかし。 なけき をおふつも
りにや いとあづしく
なりゆき て もの心ぼそげに 思ひ さとがち なるをいよ
／＼
あかす あはれなるものに おもほして。人 のそし
りをも はゞからせ給 はず。世のためし と もなり
ぬべき御もてなしなり。 かんだちめ 。 うへ人 なども。
あいなく めをそばめ つゝ いとま は ゆき人の 御 覚え
かな もろこしにもかゝることのをごりにこそ 世 は
みだれあし きこともいてきけれともてなやむほとに。
やう／＼あめのした
のあつかひくさに になりて楊貴
妃のためしもひきいでつべうなりゆくに。いと
はしたなきこと おほかれと かたじけなき 御心 さし の
た く ひなき ひとつをなくさめにて まじらひ給ちゝ の
大納言
なくなり 給て はゝ北のかた なん いにしへ の よしある
人にておやうちくしさしあたりて世のおほえはな
やか なる御かた／"＼にもいたう おとらず。何 ことの
ぎしき
るにておやうち く しさしあたりて世の覚え花
をももてなし給けれど。とりたてゝはか／"＼し
きうしろみしなければ。ことある時はなをより

Version 別國: 別本 國冬本

n

いづれの御時にか女御更衣あまたさふらひ 給
なかに。いとやむごとなききはにはあらぬが。
すぐれてときめき給ふ ありけり はじめ よりわれ
はと思ひあがり給 へる 御 かた／"＼ は めざましきもの
に おとしめ そねみ給 おなしほと 。それより
下らうの更衣たち なとまて やすからず あさゆふ
のみやづかへに つけても 。人のこゝろを うごかし。 う
らみ をおふつもりにや ありけむ いとあづしく
なりゆき もの心ぼそげに て さとがち に なるをいよ／
＼
あはれなるものに おほゝして 。人 ／＼ のそし
りをも え はゞからせ給 はず。世のためし に もなり
ぬべき御もてなしなり。 かんだちめ 。 天上人 なども。
めをそばめ て いとま は ゆき人の 覚え
かな もろこしにもかゝることのをごりにこそ は 世 の
みだれあし くはなりけれと 。やう／＼あめのした
にもあひなう人のもてなやみくさ になりて楊貴
妃のためしもひきいでつべうなりゆくに。いと
はしたなきこと おほかれと かたじけなき 御心 さし の
た く ひなき をたのみにて まじらひ給ちゝ の 大納言
も なくなり て はゝ北のかた なんくしたまへる いにし
へ の 人のよしあるにておやなしとても世のおほえはな
やか におや／＼うちくし給える／＼にも おとらず。何
ことの ぎしき
るにておやうち く しさしあたりて世の覚え花
をももてなし給けれど。とりたてゝはか／"＼し
きうしろみしなければ。ことある時はなをより
所なく心ぼそげなりさきの世にも御ちぎりやふか
かりけん世になくきよらなるたまのおのこみこ
さへむまれ給ぬいつしかと心もとながらせ給て

図3　Versioning Machineによる『校異源氏物語』等のTEIデータの表示例



- 別國 - 別本 國冬本
- 別陽 - 別本 陽明本
- 別麥 - 別本 麥生本
- 国絵 - 国文学研究資料館 絵入り源氏

- when this witness has the same reading as the lemma, the text is highlighted in white
- when this witness has a different reading from the lemma, the text is highlighted in orange
- when there is no lemma for an apparatus entry, the text of the witness is highlighted in yellow
- when an apparatus entry does not give any intelligible reading for this witness, question marks are highlighted in red ?????

Text according to 校異

いつれの御時にか女御更衣あまたさふらひ 給ける' なかにいとやむことなきゝはにはあらぬかすくれて時めき給 ありけり' はじめ'より我はと思あかり 給へる' 御'方く'めさましきものにおとしめ'そねみ給おなしほと'それより下らうの更衣たちはまして'やすからす'あさゆふの宮つかへ'につけても'人の心をのみ'うこかし'うらみ

'をおふつもりにやありけむ'いとあつしくなりゆき'もの心ほそけ'にさとかち'なるをいよくあかす' あはれなる物におもほして'人'のそしりをもえ'はゝからせ給はす世のためしに'もなりぬへき御もてなし也かんたちめ'うへ人'なともあいなく'めをそはめつゝ'いとまは'ゆき人の御'おほえ'なりもろこしにもかゝることのおこりにこそ'世も'みたれあしかりけれと'やうくあめのしたにもあちきなう人のもてなやみくさ

'になりて楊貴妃のためしもひきいてつへくなりゆく'にいとはしたなきことおほかれと' かたしけなき' 御心はへのたくひなきをたのみにてましらひ給ちゝの' 大納言はなくなりて' はゝ北の方なんいにしへの' 人のよしあるにて' おやうちくしさしあたりて' 世のおほえはなやかなる御方くにもいたう'おとらすなにことのきしきをも'もてなし たまひけれと' とりたてゝはかくしきうしろみし'なけれは事ある時'はなを'より所なく心ほそけ也さきの世にも御ちきりや

Text according to 別國

いつれの御時にか女御更衣あまたさふらひ 給' なかにいとやむことなきゝはにはあらぬかすくれて時めき給 ありけり' はじめ'より我はと思あかり 給へる' 御'方くは'めさましきものにおとしめ'そねみ給おなしほと'それより下らうの更衣たちなとまて'やすからす'あさゆふの宮つかへ'につけても'人の心を-'うこかし'うらみ

'をおふつもりにやありけむ'いとあつしくなりゆき'もの心ほそけ'にさとかちに'なるをいよく-' あはれなる物に おほゝして'人く'のそしりをもえ'はゝからせ給はす世のためしに'もなりぬへき御もてなし也かんたちめ'天上人'なとも-'めをそはめて'いとまは'ゆき人の-'おほえ'なりもろこしにもかゝることのおこりにこそは'世の'みたれあしくはなりけれと'やうくあめのしたにもあひなう人のもてなやみくさ'になりて楊貴妃のためしもひきいてつへくなりゆく'にいとはしたなきことおほかれと' かたしけなき' 御心はへのたくひなきをたのみにてましらひ給ちゝの' 大納言もなくなりて' はゝ北の方なんいにしへの' 人のよしあるにて' おやなしとても' 世のおほえはなやかにおやくうちくし給える人くにも'おとらすなにことのきしきにも'もてなし たまふれと' とりたてゝはかくしきうしろみ'なけれは事ある時'は-'より所なく心ほそけ也さきの世にも御ちきりやふかかりけむ世になく' けうら' なるたまのをのこみこさへうま

Text according to 別陽

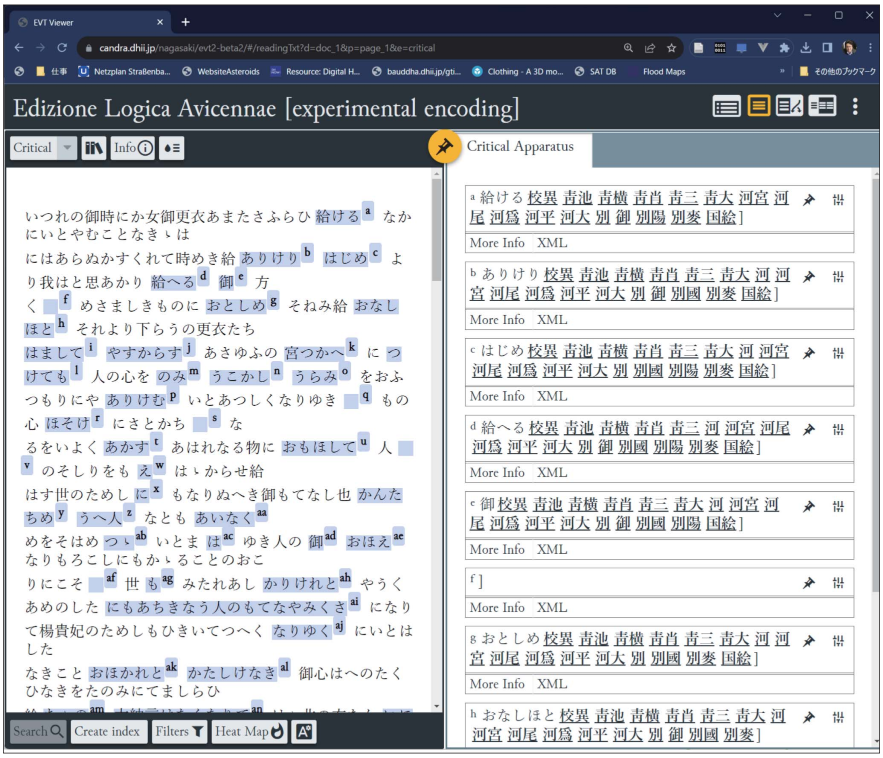
いつれの御時にか女御更衣あまたさふらひ 給ける' なかにいとやむことなきゝはにはあらぬかすくれて時めき給 おはしけり' はじめ'より我はと思あかり 給へる' 御'方くは'めさましきものにおもひをとしめ'そねみ給ける'それより下らうの更衣たちなとは'-'あさゆふの宮つかへ'につけつゝもやすからぬ事おほく思ひつむるまゝに'人の心を-'うこかし'なけき'をおふつもりにや-'いとあつしくなりゆきて'もの心ほそけ'にさとかち'なるをいよくあかす' あはれなる物に おもほして'人'のそしりをも-'はゝからせ給はす世のためしと'もなりぬへき御もてなし也かんたちめ'うへ人'なともあいなく'めをそはめつゝ'いとまは'ゆき人の御'おほえ'なりもろこしにもかゝることのおこりにこそ'世は'みたれあしきこともいてきけれともてなやむほとに'やうくあめのしたのあつかひくさに'になりて楊貴妃のためしもひきいてつへくなりゆく'にいとはしたなきことおほかれと' かたしけなき' 御心はへのたくひなきをたのみにてましらひ給ちゝの' 大納言なくなり給て' はゝ北の方なんいにしへの' よしある人にて' おやうちくしさしあたりて' 世のおほえはなやかなる御方くにもいたう'おとらすなにことのをりふしにも'もてなし たまひけれと' とりたてゝはかくしきうしろみも'なけれは事とある時'はなを'より所なく心ほそけ也さきの世にも御ちきりやふかかりけむ世になく' きよら' なる

図4　Critical Apparatus Toolboxによる『校異源氏物語』等のTEIデータの表示例

のものもある）さえあれば、誰でもパソコン上で変換することでこのような形式での表示を作成できる。横書きになってしまうのがやや残念ではあるが、日本語にも対応している。

フランスの研究者によって開発されているものとしてTEI Critical Apparatus Toolbox[※7]がある（図4）。これもVersioning Machineと同様に、各対象資料のテキストデータを個別に再構築して対応箇所の表示が可能である。前者との大きな違いは、この変換をWebブラウザ上で行えることである。したがって、異文情報等をチェックしながらTEIファイルの編集作業を続けていくという使い方もやりやすく、教育用途にも使いやすいものとなっている。さらに、PDFで印刷版を作成することが可能である点も前者との大きな違いである。ただし、このPDF版作成ツールは裏側でXSLTを使ってLaTeXに変換してからPDFに変換しており、LaTeXへの変換のところが日本語に十分対応していないため、日本語テキストでのPDF版作成は今のところできていない。次期バージョンでの対応が期待されるところである。

※7　http://teicat.huma-num.fr/



いつれの御時にか女御更衣あまたさふらひ給ける[a] なかにいとやむことなきゝは

にはあらぬかすくれて時めき給ありけり[b] はじめ[c] より我はと思あかり給へる[d] 御[e] 方

く[f] めさましきものにおとしめ[g] そねみ給おなしほと[h] それより下らうの更衣たち

はまして[i] やすからす[j] あさゆふの宮つかへ[k] につけても[l] 人の心をのみ[m] うこかし[n] うらみ[o] をおふ

つもりにやありけむ[p] いとあつしくなりゆき[q] もの心ほそけ[r] にさとかち[s] な

るをいよくあかす[t] あはれなる物におもほして[u] 人[v] のそしりをもえ[w] はゝからせ給

はす世のためしに[x] もなりぬへき御もてなし也かんたちめ[y] うへ人[z] なともあいなく[aa]

めをそはめつゝ[ab] いとまは[ac] ゆき人の御[ad] おほえ[ae] なりもろこしにもかゝることのおこ

りにこそ[af] 世も[ag] みたれあしかりけれと[ah] やうくあめのしたにもあちきなう人のもてなやみくさ[ai] になり

て楊貴妃のためしもひきいてつへくなりゆく[aj] にいとはした

なきことおほかれと[ak] かたしけなき[al] 御心はへのたく

ひなきをたのみにてましらひ

a 給ける 校異 青池 青横 青肖 青三 青大 河宮 河尾 河爲 河平 河大 別 御 別陽 別麥 国絵]
More Info XML

b ありけり 校異 青池 青横 青肖 青三 青大 河 河宮 河尾 河爲 河平 河大 別 御 別國 別麥 国絵]
More Info XML

c はじめ 校異 青池 青横 青肖 青三 青大 河 河宮 河尾 河爲 河平 河大 別 別國 別陽 別麥 国絵]
More Info XML

d 給へる 校異 青池 青横 青肖 青三 河 河宮 河尾 河爲 河平 河大 別 御 別國 別陽 別麥 国絵]
More Info XML

e 御 校異 青池 青横 青肖 青三 青大 河 河宮 河尾 河爲 河平 河大 別 御 別國 別陽 国絵]
More Info XML

f]
More Info XML

g おとしめ 校異 青池 青横 青肖 青三 青大 河 河宮 河尾 河爲 河平 河大 別 別國 別麥 国絵]
More Info XML

h おなしほと 校異 青池 青横 青肖 青三 青大 河 河宮 河尾 河爲 河平 河大 別 御 別國 別麥]

図5　Edition Visualization Technologyによる『校異源氏物語』等のTEIデータの表示例

最後に、イタリアの研究者グループによって作成されているEdition Visualization Technology※8に触れておきたい（図5）。これは、Webサーバ上で動作させるもので、このプログラムの設定ファイルを調整した上でXMLデータと共にWebサーバに設置すると、見やすい形でテキスト本文と異文情報が整形されて表示されるようになっている。他にもいくつかの便利な機能が提供されている。

なお、この種の異文化情報を表示することが可能な日本語向けの縦書き対応ツールが、上記の東アジア／日本語分科会のサイトにおいても公開されており、そちらもぜひ試してみていただきたい。

以上のように、同じような機能を提供するにしても、ツールによって使い勝手がかなり異なる。それによって利用者のタイプや見せ方も変わってくる。用途や目的に応じてさまざまなツールを開発することができるのは、XMLでデータ作成することのメリットの一つだろう。他にも用途にあわせたさまざまなツールが公開されている。

XMLはさまざまなプログラミング言語で処理できるため、XSLTだけでなく、PHPで処理するものやPython、JavaScriptなど、色々なものが提供されている。最近はあまり見かけなくなったが、かつてはJavaで書かれたものも見受けられた。印刷したりPDFで整形したりする場合にはLaTeXにコンバートして、整版はそちらに任せてしまうソリューションもある。

終わりに

TEIに準拠して作成されたテキストデータは、さまざまなプログラミング言語でXMLとして処理が可能であり、それを活かして多くのツールが開発されている。ここではその一端を見てきたことになるが、これら以外にもさまざまなツールが用途に応じて開発・公開されており、色々試してみていただきたい。また、簡単なものなら自分で開発できるというのも長所の一つであり、それを支援するための解説もさまざまなサイトで提供されている。筆者らが執筆した『人文学のためのテキストデータ構築入門』（文学通信，2022年）にもそうした解説が含まれており、さらにその発展教材はWeb公開されている※9。さらなる活用に向けた環境作りも着実に進められている※10。

人文学は、印刷メディアが登場した時代にはその最先端のメディアの特性を活かして発展してきたものであり、デジタルメディアが興隆する現代においても、やはり着々とデジタルの特性を取り込みつつ発展してきている。テキストデータの構築をめぐる人文学のこのような動向にも注目しておいていただければ幸いである。（終）

※8　https://visualizationtechnology.wordpress.com/

※9　「人文学のためのテキストデータ構築入門」フォローアップサイト
https://www.dhii.jp/dh/tei/

※10　TEI研究会では、現在、毎週オンライン研究会を開催するとともに、サイトでTEI関連のリソースの集約を行いつつある。https://tei.dhii.jp/