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（ながさき） 研宣（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

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。