人文学資料へのアノテーション～Text Encoding Initiative の挑戦

永崎研宣　一般財団法人人文情報学研究所/東京大学大学院情報学環

テキスト資料を対象として様々な考察や情報の共有を行うことに関して、人文学は、長期にわたり、蓄積を重ねてきた。その過程では、様々な明示的・暗黙的なアノテーションの手法が開発され共有されている。一言で表現するなら、Text Encoding Initiative（以下、TEI）ガイドラインとは、それをデジタル媒体上に展開するためのルールである。と言っても、それだけでは何のことだかよくわからないだろう。ここでは、TEI に関して一通りの解説をおこなった上で、日本における展開の可能性について検討したい。

上記のとおり、TEI ガイドラインは、1987 年、人文学資料における様々なメタ情報を、デジタル媒体上で、より効率的に、機械可読な形で共有するためのルールである。ありきたりではあるが、まずは歴史的経緯を概観してみることがこのガイドラインの特徴や位置づけを考える上で有益であると思われるので、どのような経緯で策定されてきているのかを簡単に見てみよう。

このガイドラインの策定は、全米人文科学基金、欧州連合、アンドリュー・メロン財団、カナダ人文社会科学協議会の支援に基づき、Association of Computers in the Humanities（ACH）、Association for Computational Linguistics（ACL）、Association of Literary and Linguistic Computing（ALLC)の三つの学会によって開始された。この際に提示された方針は以下の通りである。

テクスト電子化のガイドラインへの準備

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、同じ形式でテクストの電子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 テクスト電子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、様々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ろう。
5. ガイドラインには、そのフォーマット

において新しいテク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 テクスト記述
   2 テクスト表現
   3 テクスト解釈と分析
   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. 多くのテク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。我々は、この交換を効率化するためのツールの開発を支援するよう、支援組織に働きかける。
9. 既存の機械可読なテク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、まだ電子化されていない情報の追加に関して何か要求されるということはない。

組織的に見ると明らかに欧米中心だが、早い段階では日本からの若干の参加もあったという。

TEI ガイドラインでは、「人文学における電子テキストとはどういうものであるべきか」という議論を下敷きとして、テクストの一般的な構造が規定され、それに基づき、名前空間が定義された。これを実際の電子テキストに適用するにあたっては、当初は SGML (Standard Genelalized Markup Language)を用いて全体の構造を記述しつつ、個別のアノテーションは電子テキスト本文の中にタグを埋め込むという形で始まった。当初はまだ Web もなかった時代であり、実際にデータを共有するにあたっては FTP 等でやりとりされていたと思われる。それでも、文書同士をリンクするための参照パスの記法など、現在の XML 策定の際に採り入れられたような先進的な要素も含んでいた。しかしながら、その当時の段階では、トレーニングやアプリケーション開発などの様々な面で、SGML を用いた規格はコスト的に不利な面があり、もともとあまりお金にならない人文学資料が相手ということもあり、なかなか大きく広まるということはなかった。これが大きく広まるようになったきっかけは、おそらく、XML に準拠するようにこのガイドラインが改定されてからのことと思われる。

周知の通り、XML は策定されると同時に一気に大きな広がりを見せた。自由にタグを設定できる自由さと入れ子構造を前提とすることによる処理のしやすさは、Web 技術によく適合することとあいまって、様々なアプリケーションやプログラミング言語から、関連書籍に至るまで、あっという間に XML 技術は広まっていった。この頃はまだコンピュータ関連書籍も書店で購入するのが一般的であったが、書店の平積みの本の中にどんどん XML 関連の本が増えていっていたことは強い印象として残っている。このような中、すでに 1994 年に SGML 準拠の P3（第 3 版）が出版されていた TEI ガイドラインは、XML にも対応可能とする

ことになり、2002 年には XML 対応版の P4（第 4 版）が公開された。これによるメリットは、ほとんどあらゆるコストの低減（当ガイドライン比）であった。SGML と異なり、XML はきわめて広く普及しているため、アプリケーション開発を業者に依頼したり、あるいはデータ作成を発注したりするといった場合には、XML に対応可能な業者が多く、必然的にその種のコストは下がることになる。また、人文学研究者が自分でマークアップをしたり、マークアップされた文書を何らかの方法で処理したりしようとする場合にも、XML 関連技術であれば、トレーニングのためのコースや入門書・解説書の類が世間に多く用意されており、学習コストは格段に低くなる。もちろん、XML 関連ツールも実に様々なものが流布している。参入障壁が下がったことにより、TEI/XML を採用して電子化を行おうとする研究者やプロジェクトもどんどん増えていくことになる。そのようにして、TEI/XML は P4 にして大きく広まっていくこととなった。

しかし一方で、XML を採用したが故の問題も生じている。とりわけ、データを入れ子構造にしなければならないという点は、人文学資料における散文を相手に使用とする場合は特に問題となる。そのことは当初から予見され、一部には問題視されていたが、一方で、そもそも、物理的な構造を離れた論理的な意味でのテクストとは本来は階層構造を成しているものである、という立場もあり、解釈を整理すればきちんと XML の構造に押し込むことができるという考え方もある程度あったようである。この考え方の典型として、OHCO（Ordered Hierarchy of Content Objects）が提唱されたこともあった。この考え方についての Allen Renear の解説を見てみよう。

「どうしてこうなったのだろうか？ひとつの答えは、記述的(descriptive)マークアップというアプローチのみが、「テクストとは本来何か」という正しい視点を反映しているということである。それゆえ、記述的マークアップというコンセプトはテクストのひとつのモデルということになる。そして、そのモデルは多かれ少なかれ正しい。ここでのモデルは、テクストが特定の種類の、ある特定の方法で構造化されたオブジェクト群から構成されているということを前提としている。それらのオブジェクト群とは、章、セクション、段落、タイトル、引用、方程式（equation）、例、動作、シーン、舞台での指示、スタンザ、韻文の行などであって、頁や段、印刷上の行、フォントの行送り、垂直のスペース、水平なスペースなどのようなものではない。記述的なマークアップによって示されるオブジェクトは、テクストの知的な内容に、本質的に直接関連している。それらは、基礎をなす「論理的な」オブジェクトであり、コンポーネントである。それは、伝えたい意図を実行し体系づける際にそれぞれの役割を直接に決めることができるものである。このような「内容オブジェクト」という構造的なルールは、階層的でなければならないだろう。それらは、オーバーラップすることなくひとつずつ入れ子になっているものである。そして、それらもまた明らかに、線形の秩序を持っている。もし、あるセクションが三つの段落を「含んでいる」としたら、最

初のパラグラフは二つ目に「先行し」、二つ目は三つ目に先行する。

したがって、テクストは「秩序ある内容オブジェクトの階層構造(OHCO)」であり、記述的マークアップは、それとして機能する。なぜなら、それはその階層を定めるものであり、そして、それを、システム的な処理のために明らかにし、かつ利用するからである。この説明は、情報科学において伝統的によく知られた「間接参照」と「データ抽象化」の優位性と一致している。

多くのものがこのパースペクティブからうまくおさまるように見える。一例を挙げると、様々な種類のテクストは様々な種類の内容オブジェクト（ドラマティックなテクストにおける内容オブジェクトを法的な契約書におけるそれと比べてみよう）を有しており、そして、典型的には、内容オブジェクトがあり得るパターンは、少なくとも部分的には制約されている。手紙のある部分は、ある特定の順番で登場し、詩の行は、スタンザを外れることなく、中に登場する、等々。表現の機能は、テクストの内容オブジェクトをより容易に読者に認識させるためのものである。」

やや引用が長くなってしまったが、この考え方は、TEI ガイドラインが XML を採用する以前からあったもののようであり、階層構造を必須のものとして要求する XML への移行は、OHCO を支持する者にとっては大歓迎だったことだろう。

しかし、他方で、既に少し触れたように、OHCO は必ずしも万能ではなく、そのことを早くから指摘する研究者もいた。これは主に「オーバーラップ問題」として指摘されていたが、既存のテクストが論理的にも入れ子構造になるとは限らないという点と、それに加えて、複数の観点（たとえば、言語学的観点と意味内容としての観点）からの解釈、複数の研究者からの解釈などを細かくマークアップしようとした場合、入れ子構造として成立せず、オーバーラップせざるを得ないことがある。このため、広くデータを共有しようとする動きが拡大するにつれて、様々な解釈を一つの電子テキストに記述しようとする動きが強まっていく一方で、それまで採用していた SGML と異なり、XML はオーバーラップを基本的には許容しないため、結局の所、XML に移行したことが OHCO 的な考え方の是非を問うことになったと言ってもいいかもしれない。

XML で散文テクストにマークアップしていく際にオーバーラップに対応する手法としては、まず、マイルストーン要素と呼ばれる一部のタグを空要素とするという方法がある。これは、主に、テクストの論理構造と頁・行などの物理的な構造とのオーバーラップを避けるために、頁・行などのエレメントを空要素として記述する手法として知られている。さらにこれ以外に、近年注目されつつある方法としては、スタンドオフ・マークアップという手法がある。これは、テキスト本文へのマークアップを最小限とするか、まったく行わないか、あるいは、頁・行などの物理的構造に基づくマークアップのみにしておくなどしておき、それ以外の、内容に関わるアノテーションをすべて別の箇所に記述した上で、それぞれ、本文中の特定範囲を参照するというも

のである。スタンドオフ・マークアップに関しては、処理が複雑になるため通常のXML エディタでの編集作業はやや難しく（人文系のTEIマークアップ作業者の多くはoXygenという汎用XMLエディタを用いて電子テキスト本文へのマークアップを行っている)、近年までは人文系研究者のツールとしては少し敷居が高かったが、そのような処理がスタンドアロンなコンピュータでもある程度可能になってきていることや開発手法も整備されてきたこと等から、近年、スタンドオフ・マークアップを扱う汎用ツールが開発されフリーで公開されている。とはいえ、スタンドオフ・マークアップには、様々な面でまだまだ追求の余地があり、今後の改善や、さらに別の観点からのプロダクトの登場も期待されるところである。

また、全体として、日本や東洋の資料に関しては、Unicode が広く用いられるようになったことでようやく採用が現実的になったという段階であると言え、今後、このガイドラインがどの程度日本や東洋の資料に関しても妥当なのか、あるいは、どこを改善すれば採用が可能なのか、ということについて丁寧に議論していく必要があるだろう。

最後に、TEI ガイドラインの章の一覧を通じて、人文系研究においてどのようなアノテーションが求められているのかということについて概観して、この原稿を閉じたい。なお、以下に示すものは、P5 と呼ばれる第5版であり、第4版で問題とされた様々な点を解消した、より包括的なものとなっている。

1 The TEI Infrastructure
2 The TEI Header
3 Elements Available in All TEI Documents
4 Default Text Structure
5 Representation of Non-standard Characters and Glyphs
6 Verse
7 Performance Texts
8 Transcriptions of Speech
9 Dictionaries
10 Manuscript Description
11 Representation of Primary Sources
12 Critical Apparatus
13 Names, Dates, People, and Places
14 Tables, Formulæ, Graphics and Notated Music
15 Language Corpora
16 Linking, Segmentation, and Alignment
17 Simple Analytic Mechanisms
18 Feature Structures
19 Graphs, Networks, and Trees
20 Non-hierarchical Structures
21 Certainty, Precision, and Responsibility
22 Documentation Elements
23 Using the TEI

最初の方では、TEI の全体の構造、ヘッダの付け方、すべての TEI 文書で共通に使える XML 要素、標準的なテクストの構造、などといった章が並んでいるが、第 5 章で

は、いわゆる外字に関する扱い方が解説されている。外字は必ずしも日本語や、あるいは漢字だけでの話ではなく、たとえば中世ヨーロッパの写本では様々な異体字が用いられており、研究者の観点によってはそれらの違いが重要となることがあるため、古典学者の間では特に必要とされているのである。その後には、韻文詩や脚本、話し言葉、辞書、写本、一次資料、テクスト校訂、名前や地名などの章が用意されている。特定の個別テーマに関して解説している章は、それぞれ、対応する名前空間がモジュールとして提供され、必要に応じて取捨選択できるようになっている。また、特に、人文系として特徴的と思われるのは、第 21 章である。ここでは、確実性、正確性、情報についての責任、といったことが人によって記述される可能性があるものとして用意されているのである。日本の資料を扱う人の意見がまだそれほど反映されていないという点には最大限の配慮が必要だが、それを踏まえた上で、これらは、いわば、よく用いられる人文系資料に対するアノテーションとして求められる枠組みの現段階での一つの典型であると考えていただいてもいいだろう。

以上、非常に雑駁ながら、人文系資料のアノテーションに際して国際的に広く用いられている TEI ガイドラインについて見てきた。繰り返しになるが、これをこのまま日本の資料に適用できるかどうかについては議論が必要だが、すでに 20 年以上の議論の蓄積があるこのガイドラインをまったく無視したまま日本だけで閉じた議論をすることにも無理があるだろう。筆者としても、日本でこれに関する議論が広がっていくように色々な努力をしているところだが、本ワークショップに集うみなさまの様々な観点からのご意見や、さらに、ご関心がおありであれば、ご協力もいただければ幸いである。

参考文献

Burnard, Lou and Syd Bauman (2007), P5: Guidelines for Electronic Text Encoding and Interchange, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html .(2012/07/27 閲覧)

Cummings, James (2007), "The Text Encoding Initiative and the Study of Literature", A Companion To Digital Literary Studies, 2007, pp. 451-476.

永崎研宣「デジタルアーカイブの弁証法」『情報処理学会研究報告』CH-68(2005 年 10 月), pp. 17-24.

Nagasaki, K. (2008), "A Collaboration System for the Philology of the Buddhist Study", Digital Humanities 2008, pp. 262-263.

Renear, Allen H. (2004), "Text Encoding", A Companion to Digital Humanities, Blackwell Publishing, 2004, pp. 218-239.