chapter

8

歴史データのさまざまな応用
― Text Encoding Initiative の現在 ―

永崎研宣（人文情報学研究所）

1. デジタルテクストの特徴を活かすには？

デジタルテクストがあればテクストの扱いがとても便利になります。その是非はともかくとして、直接的にせよ、間接的にせよ、何らかのかたちで利便性が向上するということは誰もが認めることでしょう。自分は紙しか見ない、活字しか読まない、という人であっても、活字での組版を業務として行っているところはもはやごく限られており、ほとんどの活字は、デジタル組版によってコンピュータが作成した字形が印字されたものであることはご存じでしょう。

印刷されて紙媒体として読まれるものがテクストであり、われわれは、時として暗黙でもあるさまざまなルールを通じて読み取ってきました。新聞には見出しがあり段組との関係でひとつの記事が構成されていること、脚注番号を見て同ページの下部に対応する脚注がなければ章末を見て、章末にもなければ文書末尾の対応する番号を見て脚注を確認すべきこと、下線や横線が引いてあればそれに何らかの意味があると考えるべきこと、ルールはさまざまであり、時として新たに開発され、特に説明もないままに理解されてしま

うものもあればわかりにくいと消えていくものもあったかもしれません。紙媒体という、ひとつの面しか持ち得ない媒体にいかにして情報を載せて伝えようとするかという営みは、一千年を超える試行錯誤の連続であったといえるでしょう。

一方、デジタルテクストは、画面上での話ではありますが、複数の面を同時に扱うことができます。そこで多くの人が考えるのは、今まで暗黙的に共有してきたルールをきちんと書き込みつつ、通常の文章はそのまま読めるようにしたい、ということです。われわれは下線を引くとき、その下線に何らかの意味を込めようとします。多くの場合は強調したいのですが、ではどのように強調したいのでしょうか。「これ以降はこの単語に注目してもらいたい」のか「このフレーズがこれまでのすべてを一言でまとめている」のか、あるいは単に「これは重要人物の略称」なのか、下線に込められる意味はさまざまであり、それを正確に伝えることは下線のみでは不可能です。デジタルテクストでは、それを別な面に記述することが可能となります。例えば以下のものを見てみましょう。

例：それを＜キーフレーズ＞別な面に記述する＜／キーフレーズ＞ことが可能となる。

この例では、本文の中にタグと呼ばれる＜　＞に囲まれた「キーフレーズ」という文字列を 2 カ所組み込んでいます。そして、このキーフレーズが終了する箇所では「＜／キーフレーズ＞」という風に、「キーフレーズ」文字列の前に「／（スラッシュ）」が入っています。これによって、このふたつのタグに囲まれた文字列「別な面に記述する」をキーフレーズであると示そうとしています。この＜　＞内のテクスト「キーフレーズ」を通常は非表示にして、必要なときには表示できるようにすることによって、ひとつの文章についての複数の文脈での記述と提示が可能になります。さらに、「キーフレーズ」というタグを検索することで、キーフレーズとして示されている文字列を抽出することもできます。ひとつの文書の中にキーフレーズが複数登場してい

てそれぞれに同様のタグを付していれば、それらをまとめてタグ検索で抽出することができて、タグを付与する効果はより高まるでしょう。

さて、ここでもうひとつ考えてみたいのは、ほかの人と文書を共有しようとする場合です。ほかの人が同じように「キーフレーズ」タグを付けた文書を共有してくれたなら、その文書もまとめてキーフレーズを探せることになり、便利であることは間違いありません。つまり、タグを付けるだけでなく、それをほかの人と共通化することで、利便性をさらに高めることができるのです。

さて、これを 100 人がそれぞれの文書で実行してみたと想像してみましょう。100 の文書から、それぞれがキーフレーズだと思った文字列が取り出されます。このことの面白さはいうまでもないでしょう。しかし一方で、100 人それぞれが考える「キーフレーズ」がまったく同じ意味合いで選び出されることは少々難しいかもしれません。ある人は、文書に多く登場するいくつかのフレーズを選ぶかもしれませんが、一方で、登場頻度は少ないものの、文書を象徴するいくつかのフレーズを選ぶ人もいるかもしれません。ほかにもいろいろな定義の可能性があるでしょう。そうすると、100 人が作成したすべての文書から「キーフレーズ」を取り出したときにそれをもう少し統一的に扱えるようにしたいと思うなら、「キーフレーズ」がどういうものかということについて認識を共有できるようにしておく必要があります。つまり、「キーフレーズ」の定義を記述し、それを共有しなければなりません。

2. TEI 登場のコンテクスト

このようにして文章の中に注釈のようなものを埋め込んだり多様な面を記述したりすることは、1980 年代後半にはすでにそれなりにできるようになっており、2018 年現在ではかなり自由かつ便利なかたちで利用可能となっています。しかしながら、この種のことは、技術的にできるだけでは十分ではありません。各自が異なるルールでこのような記述をしてしまうと、共通の

ツールで利便性を高めたり、それぞれの成果を共有したりすることが極めて難しくなってしまいます。研究としては、誰も試みたことがない新しい記述手法に取り組むことには一定の意義がありますが、そのような記述手法はほかの誰も使ったことがないので、そのように記述されたテクストデータの活用のためには新たに活用ツールも開発しなければならなくなってしまいます。新しい記述手法を誰かが開発するたびにそれに合わせた活用ツールも開発するというのでは、いつまで経っても効率化を図ることができません。これはかなり深刻な問題にもつながり得る話であり、それを回避するためには、それほど目新しくなくても、むしろ皆が共通で使える記述手法を定めたほうがよいということになります。欧米でデジタルテクストの活用に関わる研究者たちはこれに気がついて対処を始め、それがひとつの大きな流れになったのは 1987 年のことでした。

1987 年の冬、ニューヨーク州ポキプシーに集まった彼らは、長い議論の末に、ひとつの原則を共有するに至りました。これは、会議の地の名を冠し、ポキプシー原則と名付けられました。以下に引用してみましょう。

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．
   形式に関して推奨される構文を定義する．
   テク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．
   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．
   テクスト記述
   テクスト表現
   テクスト解釈と分析
   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．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，まだデジタル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．

人文学者や情報工学者、図書館司書たちによって支えられたTEI（Text Encoding Initiative）と呼ばれるこの動向は、その後、TEIガイドラインを策定するとともに、TEI協会（Consortium）を設置し、参加者による自律的で民主的な運営体制の下、ガイドラインの改良を続けていくことになります。この動きがやがてXMLの策定に影響を与え、さらにその後、TEIガイドライン自体もXMLをベースとするものに移行することになります。

3. TEIガイドラインとは

TEI協会は、一般的な意味での標準規格というものは目指さずに、あくまでもガイドラインを提示するということを当初より決めていたようです。こ

のことの興味深さは、人文学が業績刊行の手段として著書の出版にこだわるということに深く関わっているように思える点です。人文学においては、しばしば、議論を正確に展開するために、用語とその定義、そしてそれらの関係を、一般的な用法とは必ずしも一致しないかたちで厳密に定義することがあります。いうなれば、術語体系が、著書などのひとまとまりの研究業績ごとに異なっているという状況があり得るのです。もちろん、研究資料となる資料においても同様の状況があり得ます。厳密に定められた術語体系を強要するのではなく、十分に議論した結果をガイドラインとして提示して実際の用法は利用者・利用者コミュニティに委ねるという TEI の手法は、このような人文学のあり方に寄り添ったものとして捉えることができます。

現在の TEI ガイドラインは、P5 のバージョン 3.x となっており、非常に多くの XML タグ・属性などで構成されています。ガイドラインの目次を見ることでその全体像をある程度把握することができるので、以下にそれを概観してみましょう[01]。

1 The TEI Infrastructure

2 The TEI Header

3 Elements Available in All TEI Documents

4 Default Text Structure

5 Characters, Glyphs, and Writing Modes

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

第一章では TEI ガイドラインが提示する仕組みの全体像を示しており、第二章はヘッダーについての解説です。ヘッダーは、TEI が登場した際の極めて重要な要素でした。テクストファイルにはしばしば、「このデータがどういうものであるか」ということについての説明が欠けていることがあり、それをテクストファイルの中に詳細に記述しておくために TEI ガイドラインではヘッダーの記載を必須化したのです。第三章は、すべての TEI 準拠文書で使えるエレメントの説明です。この章は大変長く、通常の文書で利用するようなエレメント・属性、そしてその使い方の例が豊富に提示されています。そして第四章は、基本的なテクストの構造のいくつかのパターンを提示しています。

第五章は、書字体系や外字などが扱われており、日本語資料を扱う上で生じてくる外字もこのルールに従うことである程度うまく情報が共有できるようになっています。欧米の資料だとアルファベットだけで済むから楽だという話が聞かれることがありますが、中世の資料では字種が多様に存在し、Unicode では表現できない外字もまだ残されていることから、Medieval Unicode Font Initiative が Unicode への外字登録を目指した活動を続けている模様です [02]。Unicode への文字の登録に関しては、近年、コンピュータの

処理性能の大幅な向上にともない、古典籍・古文書などに登場する学術用途でしか使われないような文字・文字体系も積極的に登録されるようになっています。手続きとしては、まず国際標準規格である ISO/IEC 10646 への追加が承認されてから Unicode 規格もそれに追従することになっており、新しい文字の追加は、ISO/IEC の規格への登録というかたちをとることになります。カリフォルニア大学バークレー校を拠点とする Script Encoding Initiative という団体がこの動きを幅広くサポートしています。漢字の登録に関しては、IRG という漢字検討の専門グループがいったん検討した上で ISO のワーキンググループに提案するという手順を踏むことになっています。従って、漢字を登録する場合には、まずは IRG に提案しなければならないのが現状です。ただし、IRG も近年は学術用途の漢字登録に寛容になっており、文字同定や証拠資料に関する所定のルールを踏まえた上で要登録文字であると判断されれば基本的には登録されるようになっています。時間はかかるものの、Unicode に登録することによるメリットは大きく、その必要がある文字はなるべく登録しておきたいところです。

第六章以降は、韻文詩、戯曲、演説の文字起こし、辞書、手稿の書誌情報、一次資料の記述、校訂情報、と、資料の性質に合わせた詳細な記述の仕方が提示されています。とりわけ、手稿の記述の仕方には非常に力が入っており、欧米有力大学図書館の研究司書が中世写本の目録情報をデジタル化したりデジタル画像に書誌情報を付けたりする際に広く用いられています。また、校訂テクストの異文情報の記述の仕方も充実しています。

第十三章は、固有表現に関する記述の仕方であり、これはどの種類の資料にも適用可能なとても便利なルールです。その後、少し飛ばして、第十七章では言語コーパスを作成するための単語やフレーズ、文章などのさまざまな単位に対して付与すべきタグ・属性について解説されています。

第二十章では、本来階層構造をとるべき XML のデータを TEI の形式でうまく表現するためのさまざまな工夫が紹介されています。

もうひとつ大変興味深い章は第二十一章です。この章は、人文学によるルールであることを象徴する大変興味深いものです。文書内のさまざまな要素（固有名詞とその解説など）が、どれくらいあてになるのか、そして、誰に責任があるのか、ということを明示するためのXMLタグ・属性などの記述の仕方が解説されています。

4. アップデートされるTEIガイドライン

このように、TEIガイドラインの目次を見ることでTEIの大まかな概要が見えてきます。全体的な統一感をある程度目指そうとするものの、やはり個別の資料・個別の研究手法の束縛を離れることは難しく、TEIガイドラインとしては個別の事情についてそれぞれケアすることになっています。そして、人文学全体をフォローできているわけではないため、TEI協会にはメンバーの要求に応じて分科会が設置され、そこで個別の分野・手法におけるTEI拡張の可能性が検討され、場合によってはその成果がTEIガイドライン全体に反映されることがあります。近年では、書簡の分科会を通じてそれに関するタグ・属性などが登録されました。東アジア／日本語分科会も同様にして日本語資料を対象とするさまざまな分野に必要なタグ・属性などの登録を目指して作業を続けているところです。

TEIガイドラインは人文学資料を構造的にデジタル化するための包括的なガイドラインとして策定されてきている一方、実際のところ、これまでは主に西洋の文献を対象として策定されてきました。それでも、近代日本の資料であれば多くの状況に対応可能であり、対応すべき課題は振り仮名や漢文の返り点くらいのものでした。しかしながら、古典籍・古文書になると、くずし字の連綿体やヲコト点など、ガイドラインに沿うだけでは構造化が難しい資料が増えてきます。そういった事情と対応の必要性がTEI協会においても共有されてきた結果、東アジア／日本語分科会が2016年にTEI協会に設置されることとなりました。この分科会では、TEIガイドラインの翻訳・日

本語による日本語のためのテクスト構造化ガイドライン策定・日本語資料を適正に構造化するための TEI ガイドラインの改訂案提出を目指して活動しており、遠隔ビデオ会議システムを活用して世界各地の有志により作業が進められているところです。

5. TEI ガイドラインの活用事例

TEI ガイドラインの具体的な活用事例は、欧米の資料に関しては膨大に存在しており、例えばイギリス英語の 1 億語からなるコーパス、British National Corpus に採用されていたり、シェイクスピアの戯曲に関してはさまざまな版本に合わせた TEI 準拠テクストデータが各地で公開されていたりします。XML で記述されているため、それを利用した活用の幅は非常に広く、例えば、https://www.folgerdigitaltexts.org/Ham/charChart この URL で表示されているのは「どの人物がどの幕にどういう状態で登場しているのか」を On stage, Speaking, On stage（dead), Speaking (dead) で確認できるようにした表です。これは TEI ガイドラインに沿って記述した人物情報と幕の情報を組み合わせて視覚化したものです。この場合には多少のプログラミングが必要になりますが、基本的にそれほど難しいものでなく、ごく基礎的なレベルのプログラミングができれば十分に対応可能です。また、TEI ガイドライン向けに作成された表示用プログラムもさまざまに開発されており、例えば校訂テクスト（正確にいえば学術編集版）として TEI ガイドラインに準拠して作成した XML ファイルを Versioning Machine [03] というフリーソフトウエアに読み込ませると、各版を比較できるようにしたものを作成してくれます。例えば『魔術師マーリンの予言』の複数の写本を比較しつつ注釈を付けた学術編集版を TEI ガイドラインに沿って作成し、それを Versioning Machine に読み込ませるとこのように表示してくれます [04]。同じことを、源氏物語の大規模な校訂テクスト『源氏物語大成』で試してみたものの一部を見てみましょう。

源氏物語の諸写本を TEI ガイドラインに沿って記述するのは困難ですが、

それらを集めて校訂した『源氏物語大成』の場合、活字を用いており、西洋で発展した近代的な手法を援用してテクストを作成しているため、このようにしてTEIガイドラインを適用することはさほど難しくありません。日本研究が手法において西洋の影響を強く受けていることの証左と見ることもできるでしょう。なお、縦書きになっていないのは表示の問題であり、若干のプログラミングの手間を増やせば対応可能です。

5.1. 固有表現のマークアップ

さて、本章冒頭の例のようなキーワードのタグ付けをする場合についても少し例を見てみましょう。TEIガイドラインでは第十三章で解説されているものですが、これを『走れメロス』で適用してみたものが以下の例です。

```
<said who="#メロス ">
「市を <persName corresp="#ディオニス "> 暴君 </persName> の手から救うのだ。」
</said>
```

固有名詞については人物IDを、発話についてはその話者の人物IDを付与しています。これはかなり単純な事例ですが、この人物IDを使うことで話者の特徴や呼称などについて、さまざまな傾向を視覚化することができます。図1にごく単純な視覚化の例を示しました。同様のタグ付けをいろいろな作品で行うことができれば、作品間の比較研究の手掛かりとしても有用かもしれません。

5.2. パラレルコーパスのマークアップ

原文と訳文を対応付けるパラレルコーパスを作成したいという場合にはTEIガイドライン第十七章で詳細に説明されているタグが有効です。文章ごとに <s> というタグを付けつつ、それぞれの <s> に文章IDを付けておけば、そのID同士をリンクさせた対応付け情報を作成することでパラレルコーパスを生成できる元データを用いることができます。パラレルコーパスは、自動翻訳のための教師用データに用いたり、原文か訳文のどちらかを見ながら

図 1　走れメロスの視覚化の例［05］

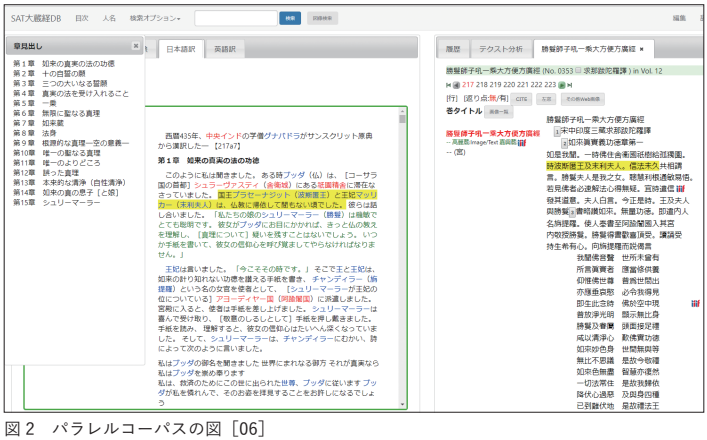

図 2　パラレルコーパスの図［06］

もう片方を一緒に閲覧したりするために用いられることが多いです。例えば図 2 の例では、大蔵経データベース上の現代日本語訳文と対応する古典中国語訳文とを並べて閲覧できるようになっています。

5.3. 校訂テクスト：学術編集版のマークアップ

ところで、TEI のコミュニティの主力を構成するグループの中には、デジタル学術編集版を作成する人たちがいます。英語では Digital Scholarly Edition などと呼ばれているものですが、いわゆる校訂テクストや校本と呼ばれる種類のテクストに近いものです。例えばキリスト教の新約聖書であれば、イエス・キリストの教えが弟子によって記録されていますが、われわれが現在読んでいるものは弟子が書いた文章そのものではありません。書写を繰り返して伝えられ、印刷技術の発展とともに印刷物として頒布され、最終的にそうして伝え残されたもの、あるいは残されたものをさらに翻訳したものをイエス・キリストの教えとして読み、理解しています。そのようにして残っていく過程では、弟子による記述が必ずしも完全に伝えられることはなく、追加や省略、修正、誤記などによって徐々にテクストは変化していきます。新約聖書というものを考えるのであれば、そうして加えられたさまざまな変化を取り除いたもとのものへとさかのぼっていこうとする通時的な方向性と、変化してきたそれぞれの時代や地域におけるテクストを再現し把握することでテクストの影響を個々の状況ごとに捉えていこうとする共時的な方向性とがあるでしょう。研究資料を紙で印刷して共有していた時代では、これらの情報を同時的に得られるようにすることは大変困難でした。しかし、デジタル媒体では、ひとつの情報群から必要な情報だけを適宜取り出して見ることができるため、(1)「それぞれの時代や地域に発見された写本の画像」(2)「その写本から文字起こししたテクストデータ」(3)「同じような（しかし時々異なっている）テクスト同士の対応付け情報」(4)「それらから判断されたよりオリジナルに近いテクスト」といったものを情報として提供し、読者・利用者が見たい部分だけをその都度瞬時に取り出して閲覧することが可能となりました。これをデジタルで共有しやすいかたちで記述するルールを TEI ガイドラインは提供しています。例えば、同様に写本で伝承されてきている源氏物語について見てみましょう（図 3）。

図 3　源氏物語の脚注の画像

このように、『源氏物語大成』では、どの行にどの写本ではどういう異文が存在するか、ということを脚注などで示しています。写本の系統を 3 つに分けているため、利用者としては、すべてを一度にまとめられるのに比べると若干使いやすいことでしょう。これを TEI ガイドラインのルールに従って記述してみたものの一部が図 4 です。

```
<app>
  <lem wit="#大成 #別麥 #別 #青池 #青横 #青三 #青肖 #別陽 #青大">給ける</lem>
  <rdg wit="#河">たまふ</rdg>
  <rdg wit="#別國">給</rdg>
</app>
```

図 4　源氏物語の critical apparatus（校異情報）記述の一例

この資料の場合、作成にあたって利用した写本が多いために記述がややごちゃごちゃしてしまっていますが、この種のごちゃごちゃした状態は、プログラミングによって解消できる面も大きいため、いきなり手作業ですべて入力しようとするのでなく、自動化できるところとできないところを見極めて、自動化できる部分はプログラミングを覚えて自分で対応してみるか、得意な

人に助けてもらうといったことを試みることをおすすめします。このようにして作成したデータは、先述の Versioning Machine を適用すると図5のように表示することができます。上記のような各資料についての校異情報から各資料のテクストを再構成した上で、カーソルを合わせるとそれぞれの資料の対応箇所に黄色いマーカーを付けてくれるようになっています。

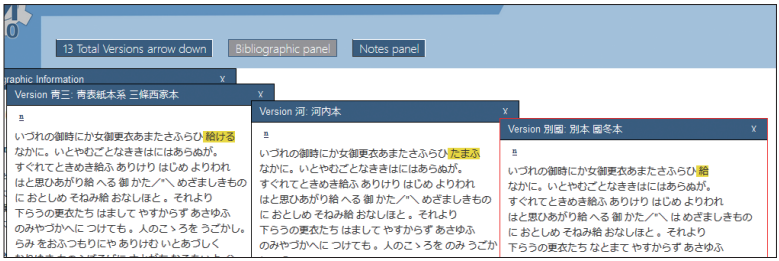

図5　Versioning Machine による表示の例

日本語資料のことをまったく意識していないメリーランド大学のプロジェクトが作成したこのフリーソフトウエアでさえ、TEI ガイドラインに準拠してデータ作成するだけでここまでのことができるのであり、また、フリーソフトウエアであるがゆえに自ら改良して縦書きなどに対応させることもできます。このようにして、国際的なデジタル・ヒューマニティーズの大きな流れに力を借りることができる上にそこにさらにフィードバックをしていくこともできるという点もまた、TEI ガイドラインのひとつの大きなメリットです。

5.4. 貨幣のマークアップ

TEI ガイドラインは、完全にそれに依拠したものしか許容しないわけではなく、むしろ、資料の特殊性に応じて拡張したり、さまざまな規格の一部として利用されたりすることも想定されています。例えば、以下のような貨幣のデータベースにおいてメタデータを記述するのに用いられている XML ベースの記述ルール（スキーマ）Numismatic Description Schema（NUDS）[07] は、貨幣の記述を目的としつつ、TEI をはじめとするいくつかのスキーマを組み合わせて構成されており、詳細情報やテクストを記述する際には TEI など

のほかの記述ルールを導入することも許容されています。

http://numismatics.org/collection/1944.100.26728

（CC BY-NC ライセンスのため全体画像の引用はしない）

このサイトでは、メタデータの出力形式として NUDS/XML 以外に、RDF/XML、TTL、JSON-LD、Linked.art JSON-LD、KML、GeoJSON、IIIF Manifest という計 8 種類のデータ形式を用意しており、連携のしやすさにも配慮している点にも注目しておきたいところです。なお、これらのデータ形式についてはほかの章で扱っているものもあるので参照してください。

5.5. 書誌情報のマークアップ

書誌情報は、ISBN を持っているような現代的な図書資料であれば、わざわざ TEI などを考える必要はないかもしれません。しかし、古典籍のような希少性の高い資料の場合には、大きさ・紙料・保存状態・来歴情報など、固有のさまざまな情報を付与しておくことが有用になります。国文学研究資料館では古典籍を調査する際に調査カードとして 31 項目の情報を記述できるようにしており、そのデータベースも公開されています。こういった情報も、なるべくコンピュータが取り出しやすいかたちに構造化されていれば可用性が高まり、利用者・読者にとっても便利です。また、一定の量が集まれば視覚化してコレクションの傾向を調べたり古典籍の流通の状況を確認したりすることもできるようになるでしょう。TEI ガイドラインでは古典籍の書誌情報を記述するためのさまざまなルールを提供しています。これに準拠して書誌情報を作成しているプロジェクトや機関は世界各地にあるようです。そのような中で、例えばケンブリッジ大学図書館は日本の古典籍の書誌情報をも TEI で公開しているので参照してください[08]。

一方、現代的な図書資料であっても、例えば青空文庫のように既存の紙の本をデジタルテクスト化した場合には、もとの紙の本の書誌情報以外に、入力・校正など、これに関わった人についての情報を記載しておきたい場合もあるでしょう。例えば、オックスフォード大学ボドリアン図書館で公開して

いるシェイクスピア作品のTEI準拠テクストでは、紙の本に関わった人の名前だけでなくデジタル化に携わった人たちの名前もその作業内容とともにヘッダーの部分に列挙されています[09]。

誰が何にどう関わったか、ということは、文化を楽しみ継承していく上で重要な要素であり、TEIがこの側面に丁寧に対応していることは、TEIの性格を端的に表しているといえるでしょう。

5.6. 画像アノテーション：IIIFとの関係

TEIはテクストデータの記述ルールから始まったものでしたが、デジタル画像の普及にともない、画像とテクストをリンクしたり、画像に対するアノテーションを記述するといったルールも導入されました。これを活用するためのツールもいくつか開発され[10]、主に研究プロジェクトにおいて活用されてきたようです。一方、画像へのアノテーションはOpen Annotationという Web上のオブジェクトに自由に注釈を付けようとする流れに淵源を持つIIIF（International Image Interoperability Framework）が2011年から欧米の有力な文化機関のITエンジニアを中心に開始され、主に文化機関がデジタルコレクションを公開する際に採用するようになりました。結果として、公開者側はIIIF対応で画像・メタデータを公開し、それを利用する側はただ閲覧するだけでなく、その任意の部分を自由に取り込んだり加工してその成果を動的に共有できるようにするといったさまざまな利活用手法の開発が世界中で取り組まれるという新しい流れが形成されています。

一方、すでに欧米の人文系研究者や文化機関はTEIに準拠した書誌情報やテクストデータを大量に蓄積してきています。そこで、IIIFで公開される画像にTEIでの蓄積をどのようにリンクさせるかという課題への取り組みが行われました。もともとTEIが持っていた画像とテクストをリンクさせる仕組みをほぼそのままIIIFに変換することが可能であったため、変換のためのプログラムはすでにいくつか実装されており、その手法も共有されつつあります。現時点ではIIIFはどちらかといえば公開されたデータを共有・

活用する仕組みという志向が強く、専門に特化したデータを作成するにはあまり向いていないため、注釈や異文情報などを埋め込んだテクストデータなどの人文学向けの基礎的なデータを作成する場合には、TEI に準拠したデータを保存用として作成し、それを IIIF に変換するというのがデータの継承性という点では安全な方法でしょう。

6. マークアップの深さをどう考えるか

TEI ではあれもできてこれもできて……という話が続くと、とりあえずテクストデータを安定して提供・共有したい場合はどういう風にすればいいのか、とか、そんなに深い構造化をするとコストがかかりすぎるから無理だ、と思ってしまうこともあるでしょう。TEI では、そういう状況に対応するべく、いくつかの解決策を用意しています。最もわかりやすいのは、TEILib（Best Practices for TEI in Libraries）[11] でしょう。これは図書館で TEI 準拠のテクストデータを作成するためのガイドラインであり、書誌情報に関しては MARC を TEI のヘッダーに変換するための対応表を提供しており、本文データに関してはマークアップの深さに関して複数のレベルを提示しています。一番浅いレベルでは OCR をかけたテクストデータをほぼそのまま利用し、もとになった画像とリンクした上で、書誌情報を記載するのみとし、レベル 2 ではレベル 1 に加えて見出しなどをマークアップすることでファイルの使いやすさを高めます。レベル 3 では、文書の基本的な構造をツリー構造になるようにマークアップしますが、パラグラフや韻文詩の行などごく基本的なマークアップにとどめます。レベル 4 では基本的な内容分析に使えるような固有表現や削除訂正などのテクストに含まれるさまざまな要素をタグ付けしますが、利用するタグは限定されます。最後のレベル 5 では、レベル 4 でも対応できない学術編集版などの深いマークアップを行うとしています。手順の自動化可能な範囲など、さまざまな情報を提示しており、現時点では英語版しかありませんが、一読の価値はあります。

7. テクストデータやツール・ノウハウを共有するには

TEI 協会では、公式 Web サイト [12] で関連プロジェクトやツールの紹介を行っており、ツールに関してはそちらを見ていただくことである程度情報が得られます。ただし、完全に網羅できているわけではないので、ほかにも Google などで探すといろいろなものを発見することができます。また、特にガイドラインに関しては GitHub 上でも公開しており、改訂のための修正案などはそちらから GitHub の仕組みを利用して提示できるようになっています。

ノウハウの共有に関しては、主にメーリングリストで質問が投げかけられるというかたちで展開した議論がアーカイビングされており、それを検索することで有用な情報をさまざまに得ることができます。

テクストデータの共有については、TEI 協会も支援する TAPAS というプロジェクトが米国で進められており、TAPAS では、TEI テクストデータリポジトリとして世界各地の TEI テクストデータのうち、ライセンス的に問題のないものが閲覧できるようになっています。

8. どうやってマークアップするか

ほかの人にも使いやすく活用しやすいテクストデータの作成ということでここまでいくつかの事例を見てきましたが、いずれも XML のタグを付けることが基本的な前提となっています。では、それをどのようにして行っていくか、ということについて以下に見てみましょう。

8.1. タグ付けルール／構造の設計

TEI でタグ付け、といわれると、何かルールが決まっていてそれに従えばいいように思ってしまいがちですが、ここまで見てきたように、分野・手法によってタグ付けの要求内容は大きく異なり、TEI に沿ってテクストデータを作ろうとする場合には、現在作ろうとしているテクストデータの目的に沿ったタグをあらかじめ選択して絞り込んでおくという作業が必要になりま

す。例えば、クラウドソーシング翻刻で有名な Transcribe Bentham というプロジェクトでは、MeidaWiki を改造し、TEI のタグのうちでこの翻刻に必要なものだけをボランティア作業者が入力するとそれに従った表示が行われるようにしています（図 6）。

ひとつのプロジェクトにおいて利用するタグを決めるプロセスにおいては、TEI ガイドラインだけではどうしても対応できないというケースへの対応も検討することになります。TEI ガイドラインを拡張するのか、ほかの XML スキーマを部分的に取り込むのか、対処方法はさまざまですが、そのような検討においては、対象資料の利用方法、あるいは少なくとも目指す利用方法をよく知っている人が主体的に関与する必要があります。問題は、文書の構造をどのように設定するかということであり、これには内容面・利用面の知識が不可欠なのです。

この種の検討においては、タグの入れ子構造などについての理解も必要になりますが、それを強力にサポートしてくれるソフトウエアもあります。商用ソフトウエアですが、汎用 XML エディタである Oxygen XML Editor を利用するのが今のところ現実的な選択肢です。Oxygen XML Editor はデフォルトで TEI 文書にも対応しており、単に XML のタグを入力しやすくしたり、作成中の文書のツリー構造を提示してくれるだけでなく、自動的に「TEI のルールに従うとその箇所で利用可能なタグ」を提案してくれる機能もあり、

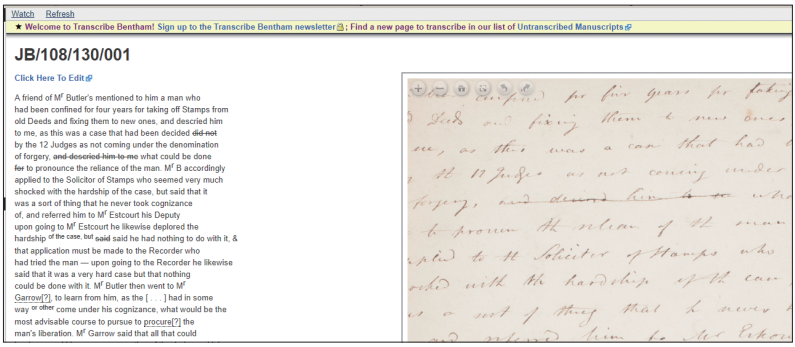

図 6　Transcribe Bentham の翻刻画面

XML に関する技術的な知識が必要な場面や面倒な作業の多くを自動的に処理してくれます。従って、どういうタグを用いてどういう構造のテクストデータを作成するかを検討する際には有用性が高いです。

8.2. どうやってマークアップするか：実際の作業

タグの付け方を決めることができたとして、次に実際のタグ付け作業についても検討してみましょう。青空文庫でも独自のタグ付けルール[13]を利用しており、TEILib では構造上はレベル 2 に相当する比較的簡便でわかりやすいものですが、これでもやはり若干のハードルを感じる向きもあるようです。いずれにしても、タグを付けるという作業に抵抗を感じる人は少なくありません。にもかかわらず TEI/XML のテクストデータが欧米で多く蓄積されてきた理由は、やはり Oxygen XML Editor の存在が大きいでしょう。このエディタを TEI 準拠モードで用いると（= TEI のスキーマを読み込ませて入力編集作業をすると）、タグの入力をするためにタグの記号を入力する必要性が少なく、入力者は、テクストを入力しながら、あるいは、入力されたテクストの構造を考えながら、タグが必要と思われる場所でエディタから提案されたタグを選びつつ作業を進めていくことができます。商用ソフトウエアであるのがなんとも残念ではありますが、インターフェイスも日本語化されており、XML だけでなく MS Office や HTML ファイル、JSON ファイル、あるいは各種プログラミング言語のファイルなど、さまざまな形式を扱えるようになっているため、一度購入すれば TEI や XML 以外でもいろいろ役立てることはできるでしょう（図 7）。

また、タグ付けを簡便にするさまざまな仕組みが用意されてきたこともあるでしょう。利用するタグを限定すれば、簡単なタグ付けシステムを用意するだけで対応できるようになります。例えば、上述の Transcribe Bentham プロジェクトでは、当初は MediaWiki を改良し、ボタンをクリックするだけで必要なタグを付与できるシステムを提供していました。また、やや汎用的なフリーソフトウエアとして、CWRC writer[14]がカナダのプロジェクトによっ

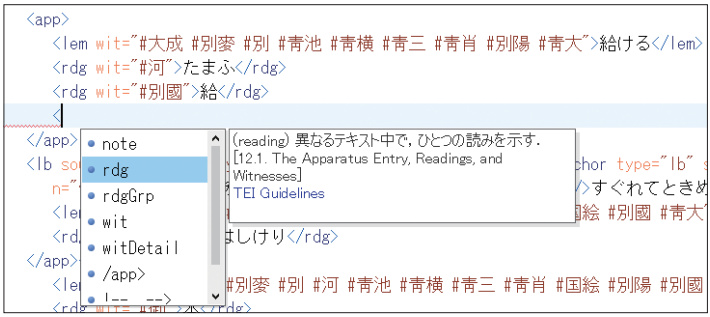

図 7　Oxygen の利用例の画面。「<」を入力するとその箇所で入力可能なタグが解説付でリストされる

て開発されており、メリーランド大学では coreBuilder [15] という TEI 準拠の外部マークアップをメニュー選択で行えるソフトウエアが開発されています。さらに別の TEI 向け汎用フリーソフトウエア開発プロジェクトも進行中です。

8.3. 自動化作業をフォローするための TEI

欧米の人文学での TEI 準拠テクストの活用と共有の仕方を見ていると、TEI 準拠ではないかたちで作成されたデータを TEI 準拠に自動的／半自動的に変換するという作業が行われることも少なくありません。TEI は、中間フォーマットとしての役割も持っており、ほかの形式から TEI に一度変換することによる多対多の膨大な変換パターンを心配することなくデータを共有できることを目指しているからです。書誌情報にせよ、本文データにせよ、何らかの構造を持っていれば、それを TEI に変換することは比較的容易です。最近は、MS-Word でさえ内部形式は XML になっているため、TEI への変換やその逆もそれほど難しいことではなくなってきています。そこで、いろいろなデータフォーマットから TEI に変換してそれを共有するという仕方も有力な選択肢となっているのです。

とはいえ、もとのデータの持っている精度を超えることは極めて難しいです。例えば、青空文庫形式のテクストデータを TEILib のレベル 2 に変換することはできますが、レベル 3 となるテクスト全体をツリー構造にするための変換は人の目と判断が必要になります。あるいは、どこに固有名詞が登場

するかということも自動的に検出することはある程度までは可能ですが、間違いや見落としが出てしまうことも多く、比較的正確に検出された信頼できるデータを作ろうとすると人の目が必要になり時間もそれなりにかかってしまいます。このような、いわば、半自動的な作業プロセスにおいても TEI は有用です。自動的にマークアップしたあと、手作業でデータを修正・整備していくにあたり、一度 TEI 準拠のデータにしておけば、データを共有しながら作業を進めていくことが比較的容易になるでしょう。例えば、MeCab で分析・注記した地名情報を含むテクストデータをあとから手で修正しようとするなら、TEI において地名を示すタグである <placeName> を付けた状態にしておけば、あとはそれを修正したり、新たに <placeName> タグを付けていったりするというワークフローが可能になります。固有名詞かどうか、文法的にはどうか、といったことに始まり、テクスト中のさまざまな側面についての注記を共有しながら進めていくことは、利用者だけでなくデータ作成者にとっても貴重な得がたい経験となることもあります。

9. おわりに

TEI は、その 30 年の歴史の中で、技術の進歩と人文学分野における方法論的内省の深化により、常に発展を続けてきています。デジタル・ヒューマニティーズ（≒人文情報学）における「方法論の共有地（Methodological Commons）」という考え方を体現する活動として、欧米ではデジタル・ヒューマニティーズの中心的な役割を果たしているもののひとつです。かつては日本語データを他言語と共通に扱うことが難しく日本での導入に意味を見出だすことが難しかった時代もあり、導入がうまくいかなかったこともあったようですが、現在は、海外で作られたさまざまなデジタルツールを日本語資料に適用することが技術的には問題がなくなってきており、あとは内容・意味の面での課題を解決すればよいという状況になっています。海外のツールやそれが依拠する枠組みを日本語資料やその研究に使えるようになるのであれ

ば、海外で進められているデジタル資料への取り組みに関する多様な観点を検討し必要に応じて適用することも可能になります。それは、単に利便性を高めることに資するだけではありません。明治の開国において西洋の人文学研究のエッセンスを取り込んで日本の人文学が成立し文化への視点が多様化したように、欧米の長い人文学の伝統から生まれ育まれてきた TEI に向き合うことで、デジタル時代のテクストのあり方への観点をより多様なものとし、日本の人文学を豊かにしていくことになるでしょう。

付記：日本語資料に TEI を適用する取り組みの現状に関しては、上述の TEI 協会東アジア／日本語分科会の活動が参考になるでしょう。https://github.com/TEI-EAJ

――注（Web ページはいずれも 2019-1-18 参照）

［01］ P5: TEI ガイドライン, http://www.tei-c.org/release/doc/tei-p5-doc/ja/html/index.html.
［02］ Medieval Unicode Font Initiative, https://folk.uib.no/hnooh/mufi/.
［03］ Versioning Machine, http://v-machine.org/.
［04］ Prophecy of Merlin, http://v-machine.org/samples/prophecy_of_merlin.html.
［05］ ここに TEI/XML を読み込ませる , https://tei-eaj.github.io/aozora_tei/tools/visualization/display_dazai.html.
［06］ SAT 大蔵経 DB, http://21dzk.l.u-tokyo.ac.jp/SAT2018/master30.php.
［07］ Numismatic Description Schema (NUDS), http://nomisma.org/nuds.
［08］ ケンブリッジ大学図書館の書誌情報の XML, https://services.cudl.lib.cam.ac.uk/v1/metadata/tei/PR-FJ-00734.
［09］ シェイクスピア作品の TEI 準拠テクストの XML, http://firstfolio.bodleian.ox.ac.uk/download/xml/F-ham.xml.
［10］ 例えば, https://mith.umd.edu/tile/. http://tapor.uvic.ca/~mholmes/image_markup/.
［11］ Best Practices for TEI in Libraries, http://www.tei-c.org/SIG/Libraries/teiinlibraries/.
［12］ TEI: Text Encoding Initiative, http://www.tei-c.org/.
［13］ 青空文庫注記形式, https://www.aozora.gr.jp/aozora-manual/index-input.html#markup.
［14］ CWRC-Writer XML editor, https://github.com/cwrc/CWRC-WriterBase.
［15］ coreBuilder, https://github.com/raffazizzi/coreBuilder.

――参考文献

» James Cummings, A world of difference: Myths and misconceptions about the TEI, Digital Scholarship in the Humanities, fqy071, 14 December 2018, https://doi.org/10.1093/llc/fqy071.
» Nancy Ide, C. Michael Sperberg-McQueen, Lou Burnard, TEI：それはどこからきたのか。そして、なぜ、今もなおここにあるのか？, デジタル・ヒューマニティーズ, 2018 年 1 巻, pp. 3-28, https://doi.org/10.24576/jadh.1.0_3.