人文学のためのテキストデータ構造化のチュートリアル

第1章

構造的なテキストデータ構築の背景

永崎研宣

version 1.0
2026.3.21 作成

本資料は、文部科学省委託事業「人文学・社会科学のＤＸ化に向けた研究開発推進事業」（JPMXP1624）において、学校法人慶應義塾が、大学共同利用機関法人人間文化研究機構から再委託を受けて作成したものです。

1．テキストデータの有用性と限界

1-1．多様なデータ形式とテキストデータの限界

デジタルメディアにおいては、テキストや数値、音声、映像、3D、時空間情報など、様々なタイプのデータを利用可能である。なかでもテキストデータは、人文学において基礎となるデータである。現在では、テキストデータ以外のデータをデジタルメディアとして様々に扱うことができるようになっているため、ここではまず、テキストデータが持つ有用性と限界について概観してみよう。

テキストデータ以外のデータ形式と比較してみると、数値データや時空間情報の場合には、より的確な計量が可能である。数値データであれば、分析対象に関して明確に数値で比較ができ、そこから統計分析も可能となる。時空間情報も、数値で示し得るという意味では数値データの一種とも言えるが、この場合には、時間にしても空間にしても、それをきちんと特定できるのであれば、前後関係や位置関係などによって対比したり定量的に検討することが可能となる。一方、画像や音声、映像、3D データ等の場合には、テキストでは表現しきれない、様々なディテールを記録し処理することが可能である。テキストはあくまでも言語表現された情報ということになるが、画像や音声、映像、3D データ等では、言語では表現できなかった部分も含めて連続した情報の全体を記録し再現することができる。たとえば、古典籍の写本資料をデジタル撮影した画像とそれを文字起こししたテキストデータとして記述した場合には、文字起こしの際には対象にならない文字の細かな癖やインクの濃淡、汚損の具体的な個々の状況など、テキストデータからは様々な情報が捨象されてしまう。音声データでも、記録された音のニュアンスのすべてを記述することは不可能だろう。映像データでは、常に移り変わる画面の隅々までもテキストデータとして記述することはやはり困難である。つまり、テキストデータは、数値データに比べると確度の高い分析をしにくく、画像や音声、映像、3D データ等に比べると情報量が少ないということになる。

1-2．捨象によって成立するテキストデータの有用性

次にテキストデータの有用性を検討してみよう。確かに、テキストは現実世界の情報をそのまま保存するメディアではない。文字起こしや記述の過程で、書字の癖、物質的状態、音声の微妙な抑揚、映像の連続的変化など、多くの情報が捨象されてしまうが、この「捨象」こそが、テキストの最大の強みともなる。テキストは、対象をそのまま再現するのではなく、**言語的・概念的に切り分け、意味づけ、再構成するための形式**である。情報量の少なさは、裏を返せば、対象を抽象化し、比較・分類・再解釈しやすい形に変換しているということでもある。数値データが数量的比較を可能にするのと同様に、テキストは意味の比較や概念の操作を可能にすると言うことができる。

1-3．学術的知識を成立させる表現形式としてのテキスト

画像や音声、映像、3D データは、非常に豊かな情報を含むが、それ自体が「何を意味するのか」を自明に示すわけではない。たとえば、古典籍の高精細画像は多くの情報を保持するが、それがどのような本文系統を示し、どのような思想的・歴史的意味を持つのかは、結局のところテキストによる記述・分析・論証を通じて示される。この点で、テキストは単なる記録媒体ではなく、解釈を外化し、共有し、批判可能な形で提示するための形式である。学術的知識が「再利用可能で、検証可能で、反論可能である」ためには、テキストによる表現が不可欠である。数値や画像は重要な根拠になり得るが、それらをどのように理解すべきかを示す枠組みは、最終的にテキストとして与えられることになる。

1-4．他形式のデータを結びつけるハブとしてのテキスト

テキストはまた、構造化や再利用の面で特異な強みを持つ。言語単位（語・文・段落）や論理単位（主張･根拠･反証）として分節化でき、注釈の付与、構造化記述、検索、引用、比較といった操作が容易である。これと対照するなら、画像や音声、映像は本質的に連続体であり、後から意味の単位を切り出す必要がある形式であると言うことができる。デジタル環境においては、この特性がさらに重要になる。テキストは、人間による読解だけでなく、計算機による処理や分析の対象にもなり得る。自然言語処理や知識グラフ、注釈付きコーパスなどは、まさに**テキストが意味を伴った離散的データで**あることによって成立している。

最後に重要な点として挙げておきたいのは、テキストが**他のデータ形式と競合するものではなく、それらを結びつけ、意味づけるハブとして機能する点**である。画像や音声、映像、3D データ等は､テキストと結びつくことで初めて､研究資源として体系的に利用可能になる。キャプション、メタデータ、注釈、解説、批評といった形で、テキストは他形式のデータを知識体系の中に位置づけることができる。

テキストは数値データほど計量的厳密性を持たず、画像や音声、映像ほどの情報量も保持できない。しかし、その制約は欠点であると同時に、意味の抽象化・解釈・共有・論証を可能にする条件でもある。テキストの有用性は、情報を「余すところなく保存する」点にあるのではなく、情報を「理解し、議論し、知識として組織化する」ための不可欠な形式であるという点にある。デジタル時代においても、むしろ多様なデータ形式が併存するからこそ、テキストは知的活動の中核としての役割を失うことはない。

2．テキストの構造とは

テキストの構造、という時には、文字が一列に並んだ一次元的・線形な構造という捉え方や自然言語としての言語的構造、内容に関する意味的・機能的構造、内容の各要素の関係についての

構造など、いくつかの構造が考えられる。本書では、テキストの構造については、単に文字列が並んでいるという状態を超えて、そのテキストがどのような単位から成り、そして、各部分がどのような役割や関係をもっているかを明示的に捉えたものとする。そして、後述する TEI ガイドラインで用いられる構造を参照し、以下の 3 つの構造という形で整理する。

1. 物理的構造
   → ページ、行、欄、写本の構成等
2. 論理的構造
   → 章・節・段落・注・引用等
3. 意味的構造
   → 人物・地名・日付・出来事・書誌要素・注釈等

これらは、人文学においてテキストを扱う際に必要とされる構造である。それぞれにオーバーラップするものであり、これらの 3 つの構造を共存させることは一筋縄ではいかない。これを人文学において利用可能な形で実現しているのが TEI ガイドラインである [1)]。ここでは、TEI ガイドラインの蓄積を背景にしつつ、この構造について検討してみよう

2-1．物理的構造と論理的構造

物理的構造、及び論理的構造のうちで簡単な要素については、人文学のどの分野でも多かれ少なかれ必要となる構造だろう。どの頁にあるテキストなのか、どの章に書かれたテキストなのか、といった情報は、引用の際に必要になるため、そういった要素を簡単に参照できることは誰にとっても有用だろう。一方、物理的構造の中でも、写本を構成する要素のようなものの場合、写本研究者にとっては有用だが、写本に書かれたテキストの内容を対象とする研究者にとっては常に必要な情報であるとは限らない。同様に、論理的構造に関しても、注や引用といった要素がそもそも必要かどうか、ということや、それをわざわざ区別して記述するといった手間をかけるべきかどうか、ということについては、研究のスタンスによって判断が異なってくるだろう。

2-2．意味的構造

意味的構造は、テキスト中に現れる語や表現を、「これは人名である」「これは地名である」「これは出来事である」といった 意味の役割 に基づいて捉える構造である。つまり、「それが何を指しているのか」「現実世界や知識世界のどの対象と対応しているのか」という点である。

2-3．意味構造の具体例

具体的な例文を通じて意味構造を見てみよう。

「天長五年、空海は京において密教の教義を説いた。」

この一文を、そのまま読むと単なる自然言語だが、意味的構造の観点から見ると、次のように分解できる。

(1) 人物（Person）

空海→ 人物（Person）

ここでは、文字列「空海」であることや主語であることではなく、「歴史上の特定の人物を指している」という意味的役割が構造の要素となる。

(2) 日付・時間（Date）

天長五年→ 日付・時間（Date）

これは、年号表現や文の冒頭にある、という論理的位置ではなく、時間的に特定可能な歴史時点という点が意味的構造の要素となる。

(3) 地名／場所（Place）

京→ 地名／場所（Place）

これも単なる名詞ではなく「特定の地理的・歴史的場所」という意味を持つ要素である。

(4) 出来事／行為（Event / Action）

密教の教義を説いた→ 出来事／行為（Event / Action）

ここでは、動詞句であったり、文末にあるといった点ではなく、「誰が」「いつ」「どこで」「何をしたか」という出来事の中核として捉えられることになる。

(5) 関係性（暗黙の意味構造）

この一文全体には、次のような関係が含まれている。

「人物（空海）が時点（天長五年）に場所（京）で出来事（説法）を行った」

これは、表面上は一文だが、意味的には 出来事を中心とした関係ネットワークになっている。

2-4．書誌情報における意味的構造の例

意味的構造は、本文だけでなく、書誌情報でも重要である。

「『金剛頂経』、唐代成立、写本一巻」

という情報があったとき、ここから抽出される意味的要素は：

『金剛頂経』→ 書名（Work）
唐代 → 成立時代（Date / Period）
写本 → 資料種別（Manifestation）
一巻 → 物理的単位（Extent）

ということになる。これらは、「「この資料が何であるか」を規定する意味的情報」ということになる。

2-5．注釈・注解における意味的構造

たとえば、本文に対する注釈で、「「京」とは、当時の平安京を指す。」と書かれている場合、本文における語「京」と注釈の説明は、意味的には、同一の場所概念を指しているという関係にある。この場合、意味的構造は、「本文と注」「異なる箇所」「異なる表現」を 同一の意味対象に結びつける 役割を果たしている。

2-6．意味的構造についてのまとめ

このように、意味的構造とは、テキスト中の表現を、人物・場所・時間・出来事・書誌要素などの意味単位として捉えることであり、それによって、異なる表現や異なる箇所、異なる資料を横断的に結びつけることができる。このようにして、意味的構造は、人文学研究における解釈・比較・再利用・知識化の基盤となるのである。

3．テキストの構造を記述する意義

3-1．テキストを「読む対象」から「扱える対象」へ変える

テキスト構造を記述するとは、単に文章を書く・保存することではなく、テキストを分析・比較・再利用・共有が可能な対象として定義することを意味する。

構造を明示しないテキストは、人間が順番に読むことはできたとしても、どこが重要で、どの単位が対応していて、どの要素が同種なのか、といったことを、機械にも他者にも正確に伝えることができない。

これに対して、構造記述は、**「このテキストは、どのような単位から成り、どういう関係を持つのか」を外在化する行為**である。

3-2．解釈の前提条件を可視化する

テキストを解釈するとき、人は無意識のうちに、「これは章だ」「これは注だ」「これは引用だ」「これは人名だ」と判断している。構造を記述することは、こうした判断を、**個人的・暗黙的な理解から共有可能・検証可能な形へと移行させる**ことに他ならない。

これは、**読みの透明性や解釈の再現性、学術的批判可能性を確保する**ための、きわめて重要な基盤である。

3-3．異なるテキストを比較可能にする

テキストの構造が記述されていれば、異なる版や異なる写本、異なる著者、異なる時代のテキストであっても、同じ種類の要素（章・注・人名など）や同じレベルの単位を対応づけて比較することが可能となる。

これは、人文学研究において不可欠な**「比較」という営為を支える技術的条件**であると言える。

3-4．テキストを知識資源へと転換する

構造を持つテキストは、検索や抽出、再構成、他のデータとの連携が効果的にできるようになる。つまり、構造記述は、**テキストを「知識の蓄積単位」に変換する操作**なのである。そして、その記述手法が国際的に共有されている場合には、国際共同研究においても極めて有効なものとなり、**国際的な知識資源が蓄積・共有**されていくことになる。

4．TEI ガイドラインに準拠した記述の意義

本書では、このテキストの構造を具体的に記述するための手法として TEI ガイドラインを主に扱う。ここでは、上記の検討を踏まえた上で、TEI ガイドラインに準拠してテキストデータの構造を記述する意義を整理してみよう。

4-1．構造記述の語彙を「私的な判断」から解放する

TEI ガイドラインを用いない場合、構造の記述は往々にして、研究者個人の判断やプロジェクト固有の慣習、暗黙の前提に依存しがちとなる。TEI を用いることの第一の意義は、このような状況に対して、**国際的に共有されたテキスト構造を記述するための語彙と枠組みを採用する**点にある。すなわち、「これは段落か」「これは注か」「これは人名か」といった判断を、TEI ガイドラインに準拠した形で明示できる。

4-2．物理・論理・意味の構造を分けて記述できる

TEI ガイドラインの大きな特徴は、**物理的構造と論理的構造、意味的構造を、混同せずに同時**

に記述できる点にある。たとえば、ページ・行という物理構造を保ったままで章・節という論理構造を表し、さらに人名・地名・出来事といった意味的構造を重ねるといったことが可能である。これは、「どの構造を、どの目的で使っているのか」を明確にし、後からの再利用や分析を容易にする。

4-3．解釈と事実記述を区別したまま記述できる

TEI ガイドラインは、原文に忠実な記述と編集者・研究者による解釈や注釈を、構造的に区別して表現できる。その結果、どこまでが資料に基づく情報か、あるいは、どこからが解釈・推定なのかが明示され、**学術的責任の所在が構造として可視化される**という利点がある。

4-4．長期的・横断的利用を前提にできる

テキストデータのみならず人文学における研究データ全般にも通じることだが、研究のために作成されたデータは、何らかの理由で使えなくなってしまう場合がある。それは、作成したプロジェクトが終了したり、担当者が異動・退職したり、対応する専用ソフトウェアが更新できなくなったりするなど、様々な理由があり得る。TEI ガイドラインは、そのような状況に対応するために策定されたものである。すなわち、特定の表示形式やソフトウェアに依存せず、その記述手法は改訂の経緯も含めてオープンアクセスの形で公開されている。そのため、時代やプロジェクト、分野を越えてテキストを再利用することが可能になる。これは、TEI が「今読むためだけの形式」ではなく、**将来の未知の利用に耐える知識基盤として設計されている**ことを意味している。

4-5．構造記述の意味についてのまとめ

テキスト構造を記述することは、テキストを解釈・比較・知識化可能な対象に変える行為である。構造は、読みの前提を可視化し、学術的共有を可能にする。そして、その記述に際して TEI ガイドラインを用いる場合には、構造記述を国際的に共有可能な語彙で行うことができ、物理・論理・意味の構造を区別しつつ重ねて記述でき、そして、解釈の透明性と長期的再利用性を同時に確保できることになる。この意味で、TEI ガイドラインを用いたテキスト構造記述は、単なる「形式化」ではなく、人文学研究そのものの成立条件を明示する行為だと言うことができる。

5．取組みの歴史

テキストデータの構造化のこれまでの流れに関して、まずは国際的状況をみてみよう。

5-1．国際編

①前提条件

1980 年代前半には IBM PC がリリースされるなどして、コンピュータでテキストを扱う環境が徐々にコンピュータ科学者以外にも広がっていくようになり、人文学のためのテキストデータ作成も徐々に普及していった。そのようななかで、人文学研究の現場では、いくつかの問題が発生していた。それは、テキストが増えすぎて人間の精読だけでは把握できない、複数の版･写本･資料を扱えるようになったものの対応関係を管理できない、電子化は進んだが全文検索しかできない、プロジェクトごとにデータを作っても他人のデータが使えず、他人にデータを使ってもらうこともできない、といったものであった。つまり、「テキストデータはあるが、研究に十分に使えない」という困難が広がっていた。

② SGML の登場：テキスト構造の「記述」

こうした状況の中で 1980 年代に登場したのが SGML（Standard Generalized Markup Language）というマークアップ言語であった。SGML は、表示（フォントや組版）ではなく**テキストの論理構造・意味構造を記述する**という点で、当時としては**革命的な思想**を持っていた。人文学の立場から見ても、章・節・注・引用や、文書の階層構造を**形式的に表現できる**という点で、理論的には非常に魅力的なものであった。

③ TEI の登場

こうした状況と SGML の登場を背景として 1987 年に始まったのが **Text Encoding Initiative（TEI）**であった。TEI が目指したのは､新しい理論を提示することではなく､**人文学で何を「構造」として共有すべきかを整理すること**であった。TEI は、章・注・人名・地名・書誌といった要素について、**共通語彙と記述指針を与える**ことで、**「他人のデータを、最低限読める状態にする」**ことを目標とした。

④ TEI/SGML 期

SGML 上に構築された TEI は、**概念整理と合意形成には成功**したが、実装の重さ・検索環境の未成熟という問題を抱え続けた。これには、コンピュータ環境がそれほど整備されていなかったこともさることながら、ツールが高価かつ複雑で、専門家でなければ扱えないという、SGML 自体の実務上の困難さも原因となった。この時期の TEI は、**「思想としては正しいが、日常的研究実務には乗りにくい」**という状況であり、十分に広まることはなかったようである。

⑤ XML の登場と TEI/XML への移行

1990 年代後半に登場した XML（Extensible Markup Language）は、上記のような状況を根本

的に変えることとなった。XML は、SGML の思想を継承しつつ、実装を大幅に簡略化した。これにより、無料ツールの普及や Web 技術との親和性、XPath / XQuery による検索といったことが実現され、**構造化テキストが実務的に扱えるもの**になったのである。XML の策定には TEI コミュニティの関係者達も深く関わる形となり、TEI の経験が活かされたようである [2)]。

XML が策定された後、TEI ガイドラインも XML へと移行した。ただしここでは、テキスト構造のモデルはそのままで、XML の枠組みを用いて記述できるようにしたという点に注意しておいていただきたい。つまり、XML はツリー構造をシンタックスとしているが、TEI のテキスト構造のモデルはツリー構造ではおさまらないため、XML の機能をうまく利用してその複雑なテキスト構造の記述を実現可能としているのである [3)]。その結果、**構造語彙はすでに整っており**、それに加えて**実装だけが一気に軽くなった**という状態が生まれた。結果として、デジタル学術編集版（Digital Scholarly Edition）やコーパス構築、大規模アーカイブ等において、**TEI/XML が事実上の国際標準として急速に普及**することとなった。

⑥ TEI/XML の国際化とガイドラインの深化

TEI ガイドラインは、普及とともに、静的な規格ではなく、**コミュニティ主導で進化する枠組み**へと変化していった。ガイドラインの内容で見ていくなら、2007 年にリリースされた TEI P5 ガイドラインでは、コミュニティでの検討を踏まえて技術委員会が改訂を行なう体制になった。特に大きな改訂としては、2011 年の 12 月には写本研究の実務的要請から `<sourceDoc>` 要素に連なる一連の要素が導入されて写本の筆写面、すなわち、物理的な写本構造（ページ・行・領域）をその見た目に忠実に構造記述できるようになった。2015 年には、書簡の記述に関して、研究現場から受信者や日付、通信状況などを構造として記述したいという強い実務的要求があり、これに応える形で一連の要素群が導入され、通信という行為そのものを記述対象にすることが可能になった。2024 年には CMC（Computer-Mediated Communication）が第 9 章として新たに追加され、電子メールや電子掲示板、SNS、チャットログ等のボーンデジタルな現代的資料を本格的に扱えるようになった。これは TEI が、人文学的コミュニケーション一般の記述枠組みへと拡張したことを示すものでもある。

⑦まとめ

結論として、TEI の歴史とは、理論と実務の緊張関係の中で、「実際に研究が前に進む水準」を探り続けてきた歴史だと言えるだろう。

5-2．国内編

①日本におけるテキストデータ構造化の形成過程

日本におけるテキストデータ構造化の歴史は、国際標準の受容史として単純に描けるものでは

なく、**人文学研究の実務、情報学的関心、学術コミュニティの制度的条件が交錯しながら、断続的に形成されてきた過程**として理解されるべきである。その初期段階として重要なのは、1990年代以降、日本国内で複数の学会・研究会を舞台に、テキストやデータを研究対象としてだけでなく、研究基盤として扱おうとする試みが並行して進められていた点である。

②国内学会における実践とその限界

国内の学会・研究会における言語資源・コーパスを中心とした動きとしては、JALLC (情報処理語学文学研究会 , Japan Association for Literary and Linguistic Computing) の活動が挙げられる。JALLC では、テキストデータを個々の研究成果の付属物としてではなく、**共有・再利用される研究資源**として整備・公開することが重視され、データ構築、配布、メタデータ記述といった実務を伴う取組みが継続的に行われてきた。ここで重要なのは、JALLC の活動が理論的議論にとどまらず、実際にデータを作り、会員相互で共有するという運用を前提としていた点である。

同時期に、人文学の側から計算機利用や電子テキストをめぐる方法論的正当性を議論する場として機能していたのが、JACH (テキスト・データベース研究会 , Japanese Association for Computers and Humanities) である。JACH では、文学・史学・哲学・仏教学・美術史など多様な分野の研究者が集まり、電子テキスト化、マークアップ、データベース化、計量的手法の導入といった実践が、研究発表や研究会を通じて蓄積された。ここでは、テキスト構造化は単なる技術導入ではなく、人文学研究として成立しうるかどうかという方法論的問いと不可分のテーマとして扱われていた。

これらと並行して、**情報知識学会**においても、テキスト、書誌、典拠、知識表現といった観点から、データ構造や記述形式をめぐる議論が進められてきた。情報知識学会では、データや知識をどのような構造で表現し、管理し、流通させるかという問題が、分野横断的に検討されており、当時、テキスト構造化はその重要な一部を成していた。また、情報処理学会 **人文科学とコンピュータ研究会**（IPSJ SIG-CH）では、より技術・実装寄りの立場から、人文学資料の電子化、マークアップ、検索・可視化、支援ツールの開発などが継続的に扱われ、テキストデータ構造化に関する具体的な試行が数多く報告されてきた。

1990 年代から 2000 年代にかけては、複数の学会・研究会を舞台に、理論だけでなく実装を伴う取組みが確かに存在していた。個別の取組みとしても、国立国語研究所による日本語コーパスの構築など、個別の大規模な事業が各地で少しずつ行なわれるようになっていた。しかし、当時のコンピュータ環境では、多様な異体字をはじめとする人文学の様々な分野において必要な様々なテキスト表現に十分に対応することができず、人文学研究者が本格的に参入するには整備不足の感が否めないという状況であった。さらに、当時の活動は、それぞれの関心や分野に強く結びついており、結果として、研究テーマの移行、世代交代、研究評価制度の制約なども相まって、これらの流れはいったん断続的・分散的なものとなり、国内の学術コミュニティの中で一貫した標準や長期的に維持される研究基盤的な枠組みとして結実するには至らなかった。

③ TEI との接続と国際標準への参与

2000 年代以降、日本の研究者が Text Encoding Initiative（TEI）に本格的に関与していく過程は、これら国内学会活動の直線的な延長というよりも、国際共同研究や海外の研究基盤との直接的な接触を通じて生じた新たな潮流として理解される。TEI は、人文学研究におけるテキスト構造を記述するための国際的ガイドラインであり、理論的完成度を競うものではなく、横断検索、再利用、長期保存といった実務的要請に応える共通語彙として発展してきた。日本の研究者にとっても、TEI は国内で合意された標準があったから採用されたのではなく、国内で断片化していた実装知や問題意識を、国際的に共有可能な形で整理し直すための枠組みとして位置づけられていった。

この文脈で大きな転換点となったのが、2016 年に TEI Consortium 内に設置された East Asian / Japanese SIG（Special Interest Group）の活動である。この SIG は、JALLC、JACH、情報知識学会、IPSJ-HC といった国内コミュニティの直接的な継承として成立したものではなく、TEI の国際的実践の場において、日本語・東アジア文献が抱える構造的課題――漢字と仮名の混在、縦書き、訓点、そしてルビ――が十分に扱われていないという認識から出発した。その結果、従来は国内の個別プロジェクトや研究会で断片的に議論されてきた問題が、**国際標準の議論の場で正式に扱われる対象**となったのである。

この SIG の成果の中でも象徴的なのが、TEI ガイドラインへの `<ruby>` 要素群の登録である。ルビは、日本語においては発音補助にとどまらず、本文と並列する別系列のテキストを極小単位で重ね合わせる仕組みであり、テキストを一次元の文字列として扱う前提を拡張する要素を含んでいる。`<ruby>` の導入は、日本語のための特例追加というよりも、日本語の書字実践が持つ構造的特性が、TEI の一般モデルを拡張する形で国際標準に反映された事例として位置づけられる。ここには、日本の TEI 実践が、単なる利用者の段階を超えて、国際標準の共同編集に関与する段階へ移行したことが明確に示されている。

④制度化の段階としての人文学 DX

さらに 2020 年代に入ると、こうした動きは国内制度の側からも再評価されるようになる。文部科学省による「人文学・社会科学の DX 化に向けた研究開発推進事業（以下、人文学 DX 事業）」では、個々の論文や著作だけでなく、テキストデータそのものを長期的に維持・再利用可能な研究基盤として整備することが明示的に求められ、構造化されたデータの重要性が政策的にも可視化された。ここに来てようやく、かつて様々な学会・研究会で断片的に実践されてきた課題が、ようやく制度的な枠組みの中で正面から扱われるようになったと言える。

以上を総合すると、日本におけるテキストデータ構造化の歴史は、国内学会・研究会による初期の実践と議論、いったんの分散と断絶、国際標準との直接的な再接続、そして制度的支援のもとでの再編成という、非連続的だが相互に無関係ではない過程として描くことができる。

JALLC、JACH、情報知識学会、情報処理学会人文科学とコンピュータ研究会等で培われた問題意識と実装経験は、共通仕様としては残らなかったものの、East Asian / Japanese SIG を通じてTEI ガイドラインへと接続され、<ruby> の導入という具体的成果として国際標準に組み込まれた。現在進行中の人文学 DX 事業は、こうした歴史を背景に、テキスト構造化を個人や個別プロジェクトの工夫にとどまらず、共同体として維持される研究基盤として位置づけ直す試みであり、日本の人文学におけるテキストデータ構造化史は新たな段階に入ったところである。すなわち、テキストデータ構造化は単なる技術導入の物語ではなく、人文学研究が自らの成果をどのような形で共有し、蓄積し、次世代へ引き継ぐかをめぐる実践の歴史としても位置づけられるのである。

注

1 これらの 3 つの構造を XML の木構造（これについては第三章を参照）として併存させることはできない。したがって、TEI ガイドラインは、XML の木構造を超えた複雑な構造を記述するためのルールを提供している。

2 これに関して特筆すべき点として、XML の関連規格の一つであり、インターネット上の情報資源における構成要素のアドレス指定を可能にする XPointer が、TEI ガイドラインにおいて先行的に整備されていた概念と実装を継承する形で成立したことが挙げられる。

3 TEI は XML に準拠しているから TEI もツリー構造だという誤解が見られる場合があるが、TEI はデータモデルとしては複数のツリー構造とグラフ構造を組み合わせたものである。XML がそうした複雑なデータモデルを記述可能であることについては、RDF (Resource Description Framework) や CIDOC-CRM といったグラフ構造のデータモデルが XML で記述可能であるということからも明らかである。