



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4章

テキストデータ 構造化とTEI

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。本資料の利用にあたっては、出典を必ず記載するなど、「文部科学省ウェブサイト利用規約」を準用（ただし、商用利用は不可とする。）してください。

1. はじめに

ここでは、第一章でも触れた TEI (Text Encoding Initiative) ガイドラインの歴史的経緯をもう少し詳細に振り返り、さらに近年の TEI コミュニティの状況について解説する。TEI はしばしば「XML でテキストをマークアップするための規格」として理解されるが、その射程は単なる記法の統一にとどまらない。TEI が対象とするのは、テキストを研究対象として再利用可能なデータにするために、どのような構造を明示し、どのような情報を記述対象とするか、という設計上の判断である。すなわち、TEI は「テキストデータ構造化」の実践を支える枠組みであり、その学習は、個々の要素名の暗記以上に、資料の性質と研究目的に応じて記述方針を選択し、それを共有可能な形で記録する能力に関わっている。

この点を理解するためには、TEI を (1) 歴史的にどのような要請から生まれたのか、(2) 誰がどのような仕組みで維持し、改訂しているのか、(3) 標準の外側でどのような実践知と教育資源が形成されてきたのか、(4) XML とスキーマといった技術的基盤がどのように「標準としての TEI」を支えているのか、という複数の側面から捉える必要がある。本章は、これらを段階的に整理することで、TEI を「仕様として理解する」視点と「方法論として理解する」視点とが、対立ではなく補完関係にあることを示すことを目的とする。

2. TEI の始まり

TEI は、人文学研究で用いられるテキストを、分野横断的かつ長期的に再利用可能な形で記述するための共通指針を整備することを目的として、1980 年代後半に始まった国際的な取り組みである。

その出発点は、1987 年に米国のニューヨーク州ポキプシーで開催された会合に遡る。当時、人文学分野では電子化テキストの利用が広がりつつあったが、分野やプロジェクトごとに記述方法がばらばらで、データの共有や再利用が困難であった。こうした状況を背景に、文学・言語学・歴史学などの研究者と情報学者が協力し、標準化されたテキスト記述法を策定する必要性が強く認識されるようになった。

この取り組みは、当初、米国の政府系人文学系研究支援機関である National Endowment for the Humanities の助成を受け、SGML (Standard Generalized Markup Language) を技術的基盤として進められた。特徴的なのは、**特定分野に限定せず、人文学全体で共有可能な抽象度の高いモデルを目指した点**である。テキストを単なる文字列ではなく、構造・意味・機能を持つ研究対象として捉え、それを明示的に記述するという発想が、TEI の中核に据えられた。そこでの合意事項は、会議開催地の名を冠し、ポキプシー原則と名付けられた¹⁾。以下に引用してみよう。

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旨とする。
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を目指す。
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。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。
 - ・テキスト記述
 - ・テキスト表現
 - ・テキスト解釈と分析
 - ・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ことはない。

1994年には最初の包括的な成果として TEI P3 Guidelines が公開され、写本・刊本・演劇テキスト・辞書・言語コーパスなど、多様な資料を記述できる枠組みが提示された。これにより TEI は、単なる技術仕様ではなく、人文学研究におけるテキスト理解の方法論を内包した国際的ガイドラインとして位置づけられるようになる。

その後、運営主体としての TEI 協会の設立（2000年）を経て、XML への本格的な移行が進められ、TEI P5 が2007年に公開された。以降、TEI は現在に至るまで継続的に改訂・拡張が重ねられている。TEI の「始まり」は、単なる技術標準の誕生ではなく、テキストをどのように構造化し、学術知として共有するかをめぐる、人文学と情報学の協働の出発点であったと言える。

3. TEIは「誰が、どのように維持しているのか」

3-1. TEI協会のガバナンス構造と意思決定

TEI (Text Encoding Initiative) は、特定の企業や国家機関によって管理されている規格ではなく、国際的な研究者コミュニティによって維持されている標準である。その運営主体が **TEI 協会 (Consortium)** である。TEI 協会は、TEI ガイドラインの策定・改訂・公開を担う非営利組織として 2000 年に設立され、以後、TEI を「一度決めたら固定される仕様」ではなく、研究や実務の変化に応じて **継続的に更新される標準**として維持してきた。

TEI 協会の組織体制は、役割分担が明確な多層構造を特徴としている。まず、全体方針や財政、運営戦略といった組織運営上の判断を担うのが **理事会 (Board of Directors)** である。理事会は、TEI 協会という組織がどのような方向性で活動していくのかを決定する役割を担っている。一方、その下で、TEI ガイドラインの内容そのもの、すなわちタグや属性、モデル設計といった技術的・概念的側面に責任を負うのが **技術委員会 (Technical Council)** である。技術委員会は、要素や属性の追加・修正、既存記述との整合性の検討、改訂提案の妥当性の審査などを通じて、TEI ガイドライン全体の一貫性を保っている。

ここで重要なのは、理事会および技術委員会の構成員が、TEI 協会の会員 (Members) による選挙・投票を通じて選出されている点である。これは、TEI が少数の権威ある専門家や特定機関によって一方的に管理される規格ではなく、**実際に TEI を利用しているコミュニティ自身によって支えられる標準であることを制度的に保証する仕組み**である。選出にあたっては、専門分野や地域的な多様性にも配慮がなされており、TEI が特定の学問分野や言語圏に偏らないよう工夫されている。

3-2. 研究と実務をつなぐ委員構成と分業体制

理事会や技術委員会のメンバーには、大学などに所属する研究者だけでなく、図書館・アーカイブ・研究インフラ機関などに雇用され、デジタル・ヒューマニティーズ (DH) の実務に携わる専門職が含まれている点も重要である。これらの専門職は、研究成果の長期保存、データ管理、公開、再利用といった実務的課題に日常的に向き合っている。そのため、TEI ガイドラインは、理論的な整合性だけでなく、**実際に運用し続けられるかどうか、将来にわたって維持できるかどうか**といった観点を含めて設計・改訂されてきた標準であると言える。

さらに、TEI ガイドラインの改訂プロセス自体が、原則としてオープンに運営されている点も押さえておく必要がある。現在、ガイドラインの修正提案や議論は、SIG (後述) や年次総会、メーリングリストで開始されることが多いが、最終段階では公開されたオンライン環境 (GitHub) 上で行われており、改訂に至るまでの議論の経過や、異なる意見がどのように検討されたのか、最終的にどの判断が採用されたのかといった過程が記録・公開されている。そのため、TEI の各

要素や設計判断について、単に「現行の仕様」を参照するだけでなく、なぜその形に落ち着いたのかという過去の改訂文脈を、誰でも遡って確認することができる。

このような改訂プロセスの透明性は、TEIをブラックボックス的な技術仕様ではなく、議論と合意形成を通じて形成されてきた標準として理解することを可能にする。学習者にとっても、TEIは「正解を覚える対象」ではなく、「どのような考え方や問題意識のもとで設計されてきたのかを読み解く対象」であることが、ここから理解できるだろう。

加えて、**Special Interest Group (SIG, 分科会)**と呼ばれる分野別・主題別の作業グループが、TEI協会の活動を実質的に下支えしている。SIGは、写本、辞書、演劇、特定の言語圏など、具体的な資料や研究領域に関心を持つ研究者・実務者によって構成され、実際の利用経験に基づく議論や改訂提案を行う。多くの場合、ガイドラインの改訂や拡張は、こうしたSIGや個別プロジェクトでの実践を出発点として提案され、技術委員会による検討を経て反映される。

以上のように、TEI協会は強い中央集権的管理を行う組織ではなく、**会員による参加型ガバナンス**、研究者と実務者の協働、そして透明性の高い改訂プロセスを通じて、TEIガイドラインを維持・更新する中枢機構として機能している。この制度的特徴こそが、次節で述べるように、TEIがコミュニティの実践と密接に結びつきながら発展してきた理由である。

4. TEIは「誰が使い、知識を生産しているのか」

4-1. 標準の外側で形成されるTEIの実践知

前節で見たように、TEIはTEI協会によって制度的に維持されているが、その標準としての実態は、協会という組織の内部だけで完結するものではない。TEIを実際に用い、運用し、その意味や可能性を具体化してきたのは、世界各地の研究者、図書館員、アーカイブ専門職、デジタル・ヒューマンティーズ(DH)の実務者、そして個別の研究プロジェクトから成る広範なコミュニティである。

TEI協会は、ガイドラインという共通の枠組みを提示しているが、TEIの具体的な使い方を一律に定めてはいない。たとえば、どの要素をどの程度まで使うか、どの情報をどの粒度で記述するかといった判断は、資料の性質や研究目的によって大きく異なる。そのため、TEIは「決められた書き方をそのまま適用する規格」ではなく、研究目的に応じて設計される記述の枠組みとして利用されてきた。

4-2. 分野別コミュニティによる実践の深化 — EpiDocの例

この点を理解するうえで有用な具体例の一つが、古代碑文・写本・パピルスなどを対象とする学術的・教育的エディションを符号化するためのガイドラインとツールを提供する、国際的かつ協働的な取り組みであるEpiDoc²⁾である。EpiDocは、碑文研究や古典学の研究者を中心とした

国際的コミュニティによって形成され、TEI ガイドラインを基盤としながら、古代資料の編集・注釈・公開に特化した詳細な利用指針や慣行を発展させてきた。ここでは、欠損の表現や補字、異読の扱いなどについて一般的な TEI ガイドラインよりも踏み込んだ記述の仕方が提示されるなど、実践に基づく合意が積み重ねられている。

EpiDoc の重要な点は、TEI 協会が上から仕様として定めたものではなく、**実際の研究プロジェクトや教育現場で生じた必要性に応じて、コミュニティ自身が TEI の使い方を具体化してきた点**にある。こうして形成された知見は、Special Interest Group (SIG) での議論や改訂提案を通じて共有され、TEI ガイドライン本体の改善にも影響を与えてきた。EpiDoc は、TEI の「応用例」であると同時に、**標準そのものを育てる役割を果たしてきたコミュニティと位置づけることができる。**

4-3. 学習と教育を支えるコミュニティ活動 — TEI by Example の例

TEI コミュニティにおける教育・学習の側面を代表する取り組みとして挙げられるのが TEI by Example³⁾ である。TEI by Example は、複数の研究者・教育者によって立ち上げられた学習支援プロジェクトであり、TEI を初めて学ぶ人々が、実際のテキストを手がかりに、その考え方と具体的な記述方法を段階的に理解できるよう設計されている。

TEI by Example の特徴は、抽象度の高いガイドラインを最初から通読するのではなく、具体的な例題と解説を通じて、「どのような判断が、どのような理由で行われているのか」を示している点にある。これにより、TEI は単なるタグ集合ではなく、**テキストをどのように捉え、どのように構造化するかという思考の枠組み**として理解されることを目指している。

TEI by Example の意義は、TEI ガイドラインという大部で抽象度の高い文書を、初学者が直接読み解くことの難しさを補う点にある。ここでは、「どの要素を使うか」だけでなく、「なぜそのように記述するのか」「別の選択肢は何か」といった判断の背景が、具体例とともに示されている。これは、TEI を単なるタグ集合としてではなく、思考の枠組みとして理解させるための教育的実践である。

重要なのは、TEI by Example もまた、TEI 協会が中央集権的に整備した公式教育プログラムではなく、コミュニティの自発的な活動として形成されてきた点である。それにもかかわらず、現在では世界各地の授業やワークショップで広く参照され、事実上の標準的学習リソースの一つとして機能している。このことは、TEI においては、教育や普及の正当性もまた、実践を通じて形成されてきたことを示している。

以上のように、EpiDoc が分野固有の研究実践を通じて TEI の使い方を深化させてきたのに対し、TEI by Example は学習と教育の観点から TEI 理解を支えてきた。両者はいずれも、TEI コミュニティがガイドラインの単なる「利用者」ではなく、その内容を実質的に形づくる共同的な知識生産の担い手であることを示す代表例である。すなわち、TEI は、中央で完成された規格が配布されるものではなく、コミュニティによる実践・教育・議論を通じて具体化され、その成果が再

び標準に還流する研究基盤なのである。

5. TEI はどこで、どのように教えられてきたのか

前節までに見てきたように、TEI は TEI 協会 によって国際標準として維持されている一方で、その教育や学習のあり方は、単一の公式カリキュラムによって一元的に規定されてきたわけではない。TEI は、大学教育、ワークショップ、研究プロジェクト、コミュニティによる教材整備など、複数の場を通じて分散的に教えられてきた。

5-1. 人文学系教育における TEI

文学・史学・文献学など、人文学を主たるバックグラウンドとする教育においては、TEI は多くの場合、具体的な資料や研究課題に即して導入されてきた。写本、文献、史料、辞書といった対象が先にあり、それをどのように記述・整理するかという実践的な問いに応答する形で、TEI ガイドラインの一部を用いた実践が教育されるのである。

このような文脈では、TEI ガイドライン全体を体系的に網羅することよりも、当該資料を扱う上で必要な要素や構造に焦点を当てる教育が行われることが多い。その結果、TEI は「共通規約」でありながら、教育の場では常にローカルな研究文脈に即して導入・解釈されてきた。

5-2. 情報学・デジタル・ヒューマニティーズを背景とする教育

一方、情報学やデジタル・ヒューマニティーズ (DH) を背景とする教育では、TEI は、仕様や標準そのものを理解する対象として扱われてきた。この場合、TEI のモジュール構造、teiHeader と本文の関係、ODD によるスキーマ定義、他のデータ記述標準との関係といった点が、教育内容として明示的に取り上げられることがある。

ただし、このような教育においても、TEI ガイドライン全体を要素レベルで網羅的に教えることは一般的ではない。むしろ、TEI という大規模な規約をどのように読み、分解し、教育目的や課題に応じて選択・再構成するかという観点が重視されてきた。したがって、情報学・DH 系の教育においても、TEI は結果として、文脈に応じて焦点化された形で扱われている。

5-3. ワークショップ・研究プロジェクトを通じた学習

大学教育と並んで、ワークショップや研究プロジェクトに付随する学習の場も、TEI 教育において重要な役割を果たしてきた。これらの場では、参加者が実際の資料を対象に作業を行いながら、どの要素を使うか、どの情報を明示的に記述するかといった判断を重ねていく。

このような学習は、教室内で完結する講義とは異なり、実践の中で理解が形成されるという特徴をもつ。その成果は、文書化され、共有され、場合によっては分野別の実践指針として整理さ

れてきた。ここでは、TEI は抽象的な知識としてではなく、具体的な作業を支える枠組みとして身につけられている。

5-4. コミュニティによる学習リソースと教育実践

TEI 教育を支えてきたもう一つの重要な要素が、TEI by Example をはじめとする、コミュニティによって整備されてきた学習リソースである。具体的な教材や事例集、オンライン学習資源は、公式教材の代替というよりも、教育現場の多様なニーズに応じて補完的に用いられてきた。

このようなリソースは、特定の教育モデルを押し付けるものではなく、授業やワークショップ、プロジェクトの目的に応じて柔軟に組み合わせられてきた。その結果、TEI 教育は、単一の「正しい教え方」に収束するのではなく、複数の教育実践が並存する形で発展してきた。

5-5. 地域差と日本における今後の可能性

なお、TEI が大学教育の中でどの程度体系的に扱われてきたかは、地域によって大きな差がある。欧米圏では、デジタル・ヒューマニティーズ教育の制度化とともに、TEI が授業やワークショップの中核的要素として継続的に教えられてきた。一方、日本を含む東アジアでは、TEI を教育課程の中で中心的に扱ってきた大学は限られており、特定の大学や研究拠点に依存した形で導入されてきたのが現状である。

しかし、この状況は、TEI 教育が日本の大学教育と本質的に相性が悪いことを意味するものではない。国内にも、TEI を用いた研究プロジェクトやデジタル基盤はすでに存在しており、近年では、デジタル・ヒューマニティーズ、研究データ管理、学術基盤整備への関心の高まりとともに、教育との接続可能性も広がりつつある。

TEI は、特定の分野に閉じた技術ではなく、テキスト資料を扱う多くの人文学分野に横断的に関わる枠組みである。そのため、必ずしも独立した専攻や大規模な教育組織を前提としなくとも、既存の人文学教育や情報教育の中に段階的に組み込むことが可能である。本書は、そのための考え方と参照点を提供することを目的としている。

5-6. まとめ：TEI 教育の全体像

以上のように、TEI の教育は、特定の場所や制度に限定されることなく、研究対象・教育目的・制度的条件に応じて多様な形で行われてきた。欧米圏と日本とでは制度的背景に差があるものの、TEI が「共通規約」でありながら文脈に応じて教えられてきたという点は共通している。

この柔軟性こそが、TEI が今後、日本の大学教育においても新たな形で展開されていく余地を持っていると考えられる。

6. SIGを通じたコミュニティ協働

6-1. SIG・コミュニティ協働によるテーマ別活動推進

国際的なデジタル・ヒューマニティーズ（DH）標準、とりわけ TEI に代表されるテキスト記述標準は、完成された固定的仕様として一方向的に提示されるものではない。むしろそれらは、多様な資料、言語、研究目的に直面する実践のなかで、継続的に解釈され、調整され、ときに拡張されながら運用されてきた。その過程を支えてきた仕組みの一つが、SIG（Special Interest Group）を中心とするコミュニティ協働によるテーマ別活動である。

SIG とは、特定の資料タイプ、方法論、あるいは言語・地域的条件に関心をもつ研究者・技術者・実務者が集まり、共通する課題を継続的に議論・共有するためのグループである。TEI のような国際標準においては、仕様策定組織としての技術委員会とは別に、SIG という比較的開かれた枠組みが設けられることで、個別プロジェクトや教育現場から生じた課題が可視化され、標準の運用や改訂に接続されてきた。SIG は、標準の例外処理を担う場というよりも、標準を多様な文脈で実効的に用いるための知見を蓄積・共有する場として機能している。どのような SIG が設置されているかを確認するには、公式サイト of SIG 一覧⁴⁾を参照されたい。

6-2. 書簡資料をめぐる協働：Correspondence SIG

SIG 活動の具体例として、書簡資料を対象とする Correspondence SIG を挙げることができる。書簡は人文学研究において重要な一次資料であるが、その構造は必ずしも単純ではない。差出人と受取人の関係、送付・受領という時間的非対称性、複数回にわたる往復の連鎖、さらには下書きや写しといった伝承形態の問題が重層的に存在する。

Correspondence SIG では、こうした特性を単一文書を前提とした一般的なテキストモデルに無理に当てはめるのではなく、TEI の枠組みの中でどのように表現しうるかを検討してきた。この SIG が導入に貢献した <correspDesc> 要素群を中心とする書簡記述の整理は、個別プロジェクトに散在していた記述方針を共有可能な形で可視化する役割を果たした。そこでは、書簡単体の構造に加え、書簡群としての関係性や、人物・場所・日付情報の記述粒度といった設計上の判断が議論の対象となった。

この SIG の活動は、書簡という資料タイプに固有の問題を通じて、エンコーディングを単なる記述作業ではなく、資料の関係構造をどのようにモデル化するかという方法論的課題として捉える視点を提供している。

6-3. 東アジア・日本語資料をめぐる協働：East Asian / Japanese SIG

地域的・言語的条件に根ざした課題を扱う事例として、East Asian / Japanese SIG の活動がある。この SIG は、日本語および漢字文化圏資料を対象とする TEI 運用上の論点を共有・検討する場

として機能してきた。欧文資料を暗黙の前提として形成されてきた TEI の枠組みを、東アジア・日本の歴史的資料に適用する際には、語や文の区切り方、注記の扱い、表記慣行に由来する情報をどの粒度で構造化するかといった設計上の検討が不可避となる。

この SIG の活動の中で特筆される成果の一つが、日本語資料に不可欠なルビ注記を扱うための <ruby> 要素群に関する貢献である。日本語テキストにおけるルビは、単なる視覚的補助ではなく、読解や意味理解に関わる情報であり、その扱いはテキストモデルの設計と密接に関係する。East Asian / Japanese SIG における議論と実践は、こうした注記を場当たりのな拡張としてではなく、TEI の既存モデルと整合的な形で位置づける方向へとつながった。

また、この SIG では、TEI ガイドラインを基盤としつつ、東アジア/日本語テキスト向けの運用ガイドラインの整備も進めている。抽象度の高い国際標準を、特定の言語文化圏で実装可能な形に整理する試みは、TEI を単に「翻訳」するのではなく、設計思想を保持したまま運用可能性を高める営みと位置づけられる。

さらに、East Asian / Japanese SIG は、継続的な普及・教育活動にも力を入れている。毎週開催されるオンライン研究会や、TEI 入門セミナーなどを通じて、初学者から実践者までを対象に、TEI の設計思想と具体的運用を段階的に学ぶ機会が提供されている (<https://tei.dhii.jp/>)。これらの活動は、SIG が専門家間の議論の場にとどまらず、標準理解の裾野を広げるための教育的インフラとしても機能していることを示している。

6-4. まとめ：SIG が果たす役割

ここでは SIG の事例として Correspondence SIG と East Asian / Japanese SIG を採り上げてみたが、対象とする資料や課題の性格こそ異なるものの、いずれも SIG という枠組みを通じて、ローカルな研究実践を共有可能な知へと転換し、標準の運用と理解を支えてきた点で共通している。SIG を通じたコミュニティ協働によるテーマ別の推進活動は、TEI を固定的な仕様として維持するための仕組みではなく、標準を多様な文脈で実効的に用い続けるための基盤となっている。

同時に SIG は、仕様がどのような議論と合意形成を経て形成・更新されてきたかを可視化する場でもある。その意味で、SIG の活動は、TEI を「方法論として理解する」教育と、「仕様として理解する」教育とを媒介する役割を果たしており、DH における標準理解を支える重要な実践の場として位置づけることができる。

7. TEI と「スキーマ」

7-1. XML におけるスキーマの役割

現在、TEI ガイドラインが主な記述形式として採用している XML は、テキストやデータを要素 (elements) と属性 (attributes) によって構造化して記述するための、汎用的な記述言語である。

XML そのものは、要素名や属性名を自由に定義できる柔軟な仕組みを持つ一方で、「どの要素を使うべきか」「どのような階層構造が妥当か」といった点については、XML 自身は何も規定しない。

このとき重要な役割を果たすのがスキーマ (schema) である。スキーマとは、ある XML 文書において、

- ・ 使用してよい要素や属性は何か
- ・ それらをどのような階層関係で配置できるか
- ・ どの要素が必須で、どの要素が任意か

といった点を、**機械的に指定・制約するための規則**である。XML 文書は、スキーマに照らして検証 (バリデーション) することで、「あらかじめ定められた構造規則に従って書かれているか」を自動的に確認できる。

これを踏まえて XML とスキーマの役割を説明すると、XML は「構造化して書くための言語」であり、スキーマは「その構造の範囲や前提を定めるための設計図」ということになる。

7-2. TEI は XML のためのスキーマを持つ標準である

Text Encoding Initiative (TEI) は、人文学資料を対象としたテキスト記述のために設計された、XML ベースの国際標準である。TEI が単なる記述慣行やタグの集まりではなく、「標準」として機能している理由の一つは、TEI としてどのような XML 文書が認められるかを定義するスキーマを、公式に策定・公開している点にある。

TEI ガイドラインでは、要素や属性の意味だけでなく、それらがどのような構造で現れるべきかについても定義されており、その定義はスキーマという形で機械可読に表現されている。これにより、異なる研究者や異なる機関で作成された TEI 文書であっても、同一のスキーマに基づいて検証・処理することが可能となる。

このことは、TEI データの**高度な共有や再利用**を可能にしている。すなわち、TEI 文書は「人間が読めば分かる」だけでなく、「機械的にも同じ前提で扱える」データとして流通し得る。その基盤となっているのが、TEI スキーマの存在である。

7-3. TEI 導入段階におけるスキーマの位置づけ

TEI を導入する初期段階では、対象とする文書の構造や、どのような情報を要素や属性として記述するのが適切かが、あらかじめ確定しているとは限らない。この段階では、TEI が提供する多様な要素や属性を試行的に用いながら、文書のどの側面を構造化するのかを検討していくことになる。

そのため、導入段階では、TEI ガイドラインに定義された要素や属性を広く含む、**包括的な TEI スキーマ**を用いることに一定の合理性がある。このようなスキーマは、記述を厳しく制限するためのものではなく、TEI という記述体系がどのような表現の可能性を持っているのかを把握

するための出発点として機能する。

この段階では、スキーマは「意識して設計する対象」というよりも、**背後で文書の妥当性を確認している仕組み**として理解されていれば十分である。TEI を書くことそのものを通じて、どの要素が頻繁に必要となり、どの表現が例外的であるかが、次第に見えてくる。

7-4. プロジェクトにおけるスキーマカスタマイズと共有

学習や作業が進み、対象とする文書の性質や記述方針が徐々に明確になってくると、スキーマに対する位置づけも変化してくる。特に、複数人が関与する研究プロジェクトや、長期的にデータを蓄積・共有する場合には、「このプロジェクトでは、TEI をどのように使うのか」という前提を、暗黙の了解ではなく、明示的に共有する必要が生じる。

このような文脈において、プロジェクト固有のスキーマを作成し、それを参加者間で共有することは、有効な手段となりうる。スキーマを通じて使用する要素や構造をあらかじめ定めておくことで、記述の揺れを抑え、データの一貫性を保ちやすくなる。また、同一のスキーマに基づいてデータを作成することで、後続の処理や分析、表示ツールをプロジェクト内で共通化しやすくなるという利点もある。ここで重要なのは、スキーマが「正解を強制する規則」ではなく、**共同作業のための前提条件を共有する装置**として機能する点である。

TEI において、このようなスキーマカスタマイズを体系的に行うために用意されている仕組みが **ODD (One Document Does it all)** である。ODD とは、TEI における要素や属性の定義、使用方針、構造制約を、人間が読むための説明と、**機械が処理するための定義とを併せ持つ形で記述する枠組み**である。TEI ガイドラインそのものも ODD で記述されており、そこから RELAX NG や W3C XML Schema といった具体的なスキーマが生成されている。

プロジェクトが独自の ODD を作成するという行為は、単に使用可能な要素を制限する技術的設定作業ではない。それは、「このプロジェクトでは、どの情報をテキストとして構造化し、どの側面を記述対象とするのか」という設計判断を、文書として明示的に宣言する行為である。ODD は、スキーマ生成のための中間形式であると同時に、プロジェクトの方法論や記述方針を記録・共有する文書として機能する。

もっとも、ODD を直接記述することは、導入段階の利用者にとって必ずしも容易ではない。そのため TEI では、ODD に基づくスキーマカスタマイズを、より現実的な形で支援するための Web アプリケーションとして TEI Roma⁵⁾ を提供している。TEI Roma は、ブラウザ上の操作によって、使用するモジュールや要素を選択し、それに基づく TEI スキーマを生成できるツールである。利用者は、ODD を一から書かなくとも、プロジェクトの目的や対象資料に応じたスキーマを段階的に構成することができる。

TEI Roma の意義は、スキーマ生成を単なる技術的作業として隠蔽するのではなく、**どの要素を採用し、どの構造を許容するかという設計判断を可視化する点**にある。Roma による選択結果は ODD として出力され、さらに具体的なスキーマへと変換されるため、生成物はそのまま検証

や共有に用いることができる。これにより、初期段階では Web アプリを用いた直感的な操作から始め、必要に応じて生成された ODD やスキーマを読み解き、調整するという段階的な関わり方が可能となる。

ただし、こうしたスキーマカスタマイズや ODD の利用は、TEI を用いるすべての場面で必須となるものではない。個人作業や学習段階においては、既存の TEI スキーマをそのまま用いるだけでも多くの場合は十分であり、独自のスキーマ設計を行う必要はない。スキーマの作成や調整は、あくまでも協働作業やプロジェクト運営の文脈において、記述方針を明示的に共有する必要が生じた場合に検討される付加的な手段と位置づけるのが適切である。

このように、TEI におけるスキーマカスタマイズと共有は、最初から厳密な設計を要求するものではなく、実践と理解の進展に応じて段階的に導入される仕組みとして構想されている。その中核に位置する ODD と、それを実務的に支援する TEI Roma は、TEI を単なる XML 規格ではなく、人文学研究の方法論と協働作業を支える柔軟な研究基盤として成立させている重要な要素である。

7-5. まとめ

本節では、XML、TEI、スキーマの関係を整理しながら、TEI におけるスキーマの位置づけを段階的に説明してきた。XML は汎用的な記述言語であり、その構造や使用可能な要素・属性を具体的に定めるのがスキーマである。TEI は、人文学資料を対象とした XML の利用方法を体系化した国際標準であり、公式に策定されたスキーマを持つことで、データの検証、共有、再利用を可能にしている。

TEI の導入段階では、対象文書の性質や記述方針が十分に整理されていない場合が多く、要素や属性を広く含む包括的な TEI スキーマを用いて、記述の可能性を試行的に探ることに意義がある。この段階では、スキーマは設計対象というよりも、背後で妥当性を支える仕組みとして理解されていけば十分である。

一方、複数人による協働作業や研究プロジェクトの文脈では、スキーマを共有することで記述方針を明確にし、作業の効率化や後続処理の容易化が図られる場合がある。ただし、スキーマのカスタマイズは必須ではなく、既存の TEI スキーマを用いるだけでも多くの作業は遂行可能である。重要なのは、スキーマが常に設計されるべきものではなく、目的や体制に応じて付加的に用いられる仕組みであるという点を理解することである。

このように、XML、TEI、スキーマ、そしてカスタマイズの関係を段階的に捉えることで、TEI を技術的な規則の集合としてではなく、人文学研究の実践を支える柔軟な基盤として位置づけることが可能となる。

8. まとめ

本章では、TEI ガイドラインの成立史（ポキプシー原則から TEI P3、TEI 協会設立、XML 化まで）を整理し、TEI が固定された仕様ではなく継続改訂される標準として発展してきたことを示した。あわせて、理事会・技術委員会・会員による選挙・公開された改訂過程（GitHub）といったガバナンス構造により、特定機関に依存しない形で維持されている点を確認した。さらに、EpiDoc や TEI by Example、SIG 活動（Correspondence SIG / East Asian / Japanese SIG）を例に、実務・教育の知見がコミュニティで蓄積され標準へ還流する循環を概観した。最後に、XML スキーマと ODD の関係を説明し、導入段階では包括的スキーマを用い、協働・長期運用ではスキーマ共有とカスタマイズが品質管理と再利用性を支える、という段階的理解を提示した。

注

- 1 Text Encoding Initiative, Design Principles for Text Encoding Guidelines, 14 December 1988, <https://tei-c.org/Vault/ED/edp01.htm>
- 2 <https://epidoc.stoa.org/>
- 3 <https://www.teibyexample.org/exist/>
- 4 <https://tei-c.org/activities/sig/>
- 5 <https://roma.tei-c.org/>