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 7 章

構造化テキストの作成

――基礎演習 3――

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．固有表現（人物・地名情報等）の記述と参照 ― 演習 1-4（第 6 章から続く）

演習 1-3 で、地名が登場するところには <placeName> のタグを付与した。これらは同じ実体を指しているにも関わらず表記が異なる場合がある。人名の場合も同様のことがある。このような場合、近年ではコンピュータでもある程度の正確さで同一かどうかを判定できるようになってきている。しかし、人による判断にはまだ精度では及ばない。一方で、コンピュータによる判断の精度を高めるためには手本となるデータが必要でもある。そのようなことから、人手で対応可能なデータ量であれば人手で作成してしまうというのも未だ有力な選択肢である。TEI ガイドラインでは、そういった情報を記述するためのメカニズムを提供しているので、ここではそれを試してみよう。

①人名情報の記述

1. まず、本文中に登場する人名に <persName> タグをつけよう。この書簡には「其許モ筆モ達者ト」のところに漱石の長女の筆子、<closer> のところに漱石と鏡子の名前がそれぞれ略された表記で登場しているので、それぞれ以下のようにタグ付けしてください。

```
其許モ<persName>筆</persName>モ達者ト
<closer><lb/>
    <signed><persName>金之助</persName></signed>
    <lb/>
    <salute><persName>鏡</persName>どの</salute></closer>
```

2. 次に、人物の情報を記述してみます。人物情報をリストする <listPerson> は、本文中の記述ではない場合にはヘッダの <sourceDesc> や <text> の中の <back> に記載できる。<back> に記載する場合、本で言うところの巻末付録のような位置づけになるが、ここではそのやり方を採ってみる。

   ① <text> エレメントの最後にある </body> に続けて <back> エレメントを挿入しよう。

   ② <back> の中に、<listPerson> を入力しよう。ここに、二人の人物情報を <person> で列挙する。まずは、<listPerson> の中に <person> エレメントを一つ入力しよう。

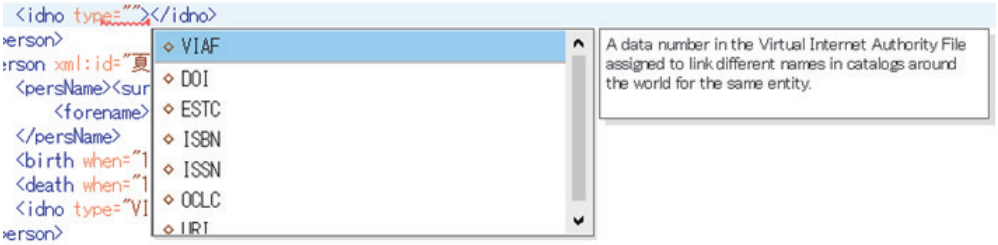
   ③ 一つ目の <person> には夏目漱石の人物情報を書こう。まず、この人物に @xml:id で識別子を与えておく。ここでは「夏目漱石」としておく。

   ④ 人名を記述する。本名とペンネームがあるので、それぞれを記述し、@type 属性で区別する。姓名はそれぞれ <surname> エレメントと <forename> エレメントでタグ付けしておく。

   ⑤ 生年月日、没年月日を記載できるようになっているので、これを記載してみよう。それ

ぞれ、<birth> と <death> というエレメントが <person> の中に書ける。ここで @when 属性で ISO8601 準拠の日付を記述し、タグ付けされるテキストとしては和暦を書いておこう。

⑥ 外部で蓄積されている関連データとつながるように、<idno> を付与しておこう。ここで以下のように、候補となる識別子名がリストされる。いずれも人文学において有用なものなので、Web 検索等でそれぞれ確認してみよう。ここでは VIAF を選択した上で、その URI を VIAF のサイトで確認して記載しておこう。

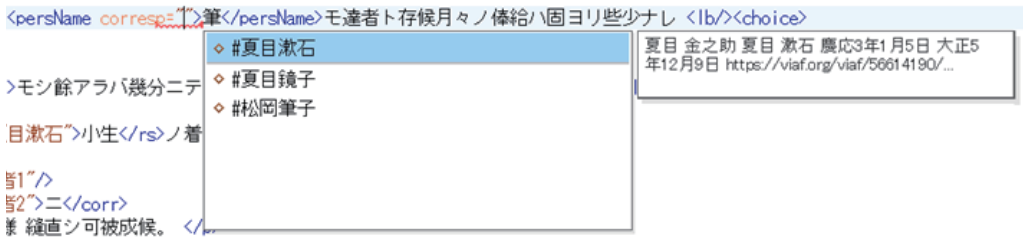


⑦ 夏目鏡子、筆子についても同様に記述しておこう。二人とも旧姓があるのでそれも記載しておこう。そうすると、<listPerson> は以下のようになっているはずである。

```
<listPerson>
    <person xml:id="夏目漱石">
        <persName type="本名"><surname>夏目</surname>
            <forename>金之助</forename>
        </persName>
        <persName type="ペンネーム"><surname>夏目</surname>
            <forename>漱石</forename>
        </persName>
        <birth when="1867-02-09">慶応3年1月5日</birth>
        <death when="1916-12-09">大正5年12月9日</death>
        <idno type="VIAF">https://viaf.org/viaf/56614190/</idno>
    </person>
    <person xml:id="夏目鏡子">
        <persName type="旧名"><surname>n中根</surname>
            <forename>鏡子</forename>
        </persName>
        <persName type="本名"><surname>夏目</surname>
            <forename>鏡子</forename>
        </persName>
        <birth when="1877-07-21">明治10年7月21日</birth>
        <death when="1963-04-18">昭和38年4月18日</death>
        <idno type="VIAF">https://viaf.org/viaf/58810142/</idno>
    </person>
    <person xml:id="松岡筆子">
        <persName type="旧名"><surname>夏目</surname>
            <forename>筆子</forename></persName>
        <persName type="本名"><surname>松岡</surname>
            <forename>筆子</forename>
        </persName>
        <birth when="1899-05-31">明治32年5月31日</birth>
        <death when="1989-07-07">平成元年7月7日</death>
    </person>
</listPerson>
```

⑧ ここでつけた @xml:id を、本文中の <persName> から参照するように記述しよう。まず、

「筆」のところで、@corresp 属性で入力しようとすると、Oxygen は以下のように入力作業を補助してくれる。



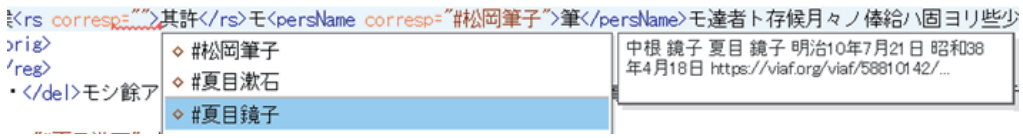
これは、@xml:id を入力すべき属性値の箇所に入力する際に、この文書中に登場する @xml:id とその内容を候補として一覧し、そこから選択するだけで入力できるようになっている。

ここでは「# 松岡筆子」を選んでおこう。ここで、@corresp 属性に書かれる ID の文字列には「#」が最初についている。この「# で始まる文字列」は、XML 文書の中で属性の値として記載された場合には、同じ文書内のその ID（ここでは @xml:id の値）を参照することを意味する。ここでは属性名が「corresp」であり、「何らかの対応を持っている」ことを意味する。このように ID を参照できる属性は TEI では他にもいくつかある。なかでも、@ref, @sameAs 等は比較的よく用いられる。

同様にして、<closer> の中の二人の <persName> にも @corresp をつけてみよう。

②代名詞の記述とリンク

より深い読解の成果を共有し、より深い分析に役立てたりするために、代名詞などの参照文字列をタグ付けし、ID 参照を与えることもできる。これには <rs> エレメントを用いる。たとえば、以下の例では、「其許」に <rs> をタグ付けし、@corresp で # 夏目鏡子の ID を付与しようとしている。



他にも「其許」「小生」といった表現が本文中にあるので、それぞれ、同様に <rs> と @corresp での ID 参照を付与してみよう。

③地名情報の記述

地名に関しては、<listPlace> と <place> を用い、人物情報と同様に、@xml:id と @corresp で参照できる。また、<correspDesc> のところで入力した方法で座標情報を記述できる。それらを踏まえて、地名情報を記述し、本文中に登場する地名から参照されるようにしてみよう。さらに、<correspDesc> に書かれている <placeName> をここの <place> を参照する形にすれば、地名情報の記述を一カ所にまとめることができ、作業効率を高めることができる。それもぜひ試していただきたい。

なお、このような状況で座標情報を記述する場合、漱石が立ち寄った場所など、特定の場所の座標を記述するのか、あるいは、自治体等の代表的な庁舎等の地点の座標にするのか、ということを決めておく必要がある。上述の <correspDesc> の <geo> では、実際の場所になるべく近いと思われる場所の座標を記述しているが、本文中での地名への言及については、むしろ、代表的な地点の座標とした方がよい場合もあるだろう。これも一貫した方針を立てることが望ましいため、適宜検討されたい。

2．人名典拠情報とその活用

2-1．人名の表記揺れと同一性の問題

上記の演習では、本文中に登場する人名に <persName> 要素を付与し、さらに <listPerson> を用いて人物情報を整理した。同様の作業は地名に対して <placeName> を用いる場合にも行っている。

ここで改めて確認しておくべき重要な点は、**同じ実体を指していても、人名や地名の表記は必ずしも一意ではない**という事実である。たとえば、

- 略称・通称・雅号・筆名
- 漢字表記の揺れ
- 旧字体・新字体
- 旧姓・婚姻後の姓
- 欧文転写の違い

といった要因により、同一人物であっても複数の表記が共存する。これは歴史資料・文学資料において特に顕著であり、単純な文字列一致によって同一人物を判定することは困難である。

近年では、文字列類似度や文脈情報を用いて、コンピュータがある程度の精度で同一性判定を行えるようになってきている。しかし、その精度は依然として人間の判断には及ばず、また、機械学習による判定精度を高めるためには、学習用の正解例データ（教師データ）が必要である。

そのため、**人手で対応可能な規模のデータについては、人間が責任を持って同一性を判断し、その結果を構造化して記述する**という方針は、現在でも極めて有力である。TEI ガイドラインは、そのような判断結果を記述し、再利用可能な形で共有するための仕組みを提供している。

2-2．人名典拠情報とは何か

人名典拠情報とは、ある人物を一意に識別するための標準化された情報の集合である。これには通常、

- 正式な名前
- 別名・異表記
- 生没年
- 職業・活動分野
- 他のデータベースとの対応関係

などが含まれる。人名典拠の目的は、「この名前が誰を指しているのか」を、人間だけでなくコンピュータにも分かる形で明示することにある。

以下では、人文学分野で特によく利用される代表的な人名典拠情報源を取り上げ、それぞれの特徴と意義、TEI との関係を整理する。

2-3．主要な人名典拠情報源

① NDL Web Authority

NDL Web Authority は、国立国会図書館が提供する典拠情報サービスであり、日本語資料に基づく人名・団体名・件名などの典拠データを公開している。

● 特徴

- 日本人作家・研究者・歴史的人物に強い
- 日本語表記の揺れや旧字体に関する情報が充実
- 国立国会図書館の書誌データと密接に結びついている

● 意義

- 日本語資料を扱う研究において、最も信頼性の高い基盤の一つ
- 国内外の他典拠（VIAF 等）へのリンクも提供されている

● TEI での活用

- `<idno type="NDL">` 等として識別子や URI を記述
- 国内研究プロジェクト間での同一人物の共有基準として有効

② VIAF

VIAF（Virtual International Authority File）は、各国の国立図書館や大規模機関が持つ典拠データを統合した、国際的な人名典拠サービスである。

● 特徴

・複数機関の典拠を統合したクラスタ構造
・多言語・多表記を一つの人物にまとめている
・URI による安定した参照が可能

● 意義

・国際的なデータ連携における事実上のハブ
・異なる言語圏の研究成果をつなぐ役割を果たす

● TEI での活用

・`<idno type="VIAF">` に VIAF URI を記述
・異なるプロジェクト間で人物同一性を保証するための参照点として利用

③ ISNI

ISNI（International Standard Name Identifier）は、人物や団体に対して付与される国際標準識別子である。

● 特徴

・ISO 標準に基づく識別子
・人物・団体を横断的に対象とする
・商業出版・学術出版の双方で利用されている

● 意義

・国際的に安定した「名前の番号」として機能
・他の典拠（VIAF 等）とも相互参照関係にある

● TEI での活用

・`<idno type="ISNI">` に識別子を記述
・長期的なデータ共有・保存を見据えた参照先として有効

④ Getty Vocabularies（ULAN）

Getty Vocabularies の一つである ULAN（Union List of Artist Names）は、美術・文化史分野の人物を中心とした典拠データベースである。

● 特徴

・芸術家・文化人に特化
・活動時期・関係性・地域情報が豊富
・Linked Open Data として公開されている

● 意義

・美術史・文化史資料との高い親和性

・人物間の関係性を含む分析に向く

● TEI での活用

・`<idno type="ULAN">` 等として URI を記述

・人物ネットワーク分析や可視化との接続が容易

⑤ Wikidata

Wikidata は、ウィキメディア財団が運営する、オープンかつ協働的に構築される知識ベースである。Wikipedia 各言語版の背後で利用される構造化データの基盤として設計されており、人名・地名・出来事・作品など、多様な実体に対して識別子（Q 番号）を付与し、属性（プロパティ）として情報を記述する仕組みを持つ。

● 特徴

・世界規模で共同編集されているオープンな知識基盤

・多言語ラベル・別名・説明文を標準で保持

・VIAF、ISNI、NDL Web Authority、Getty ULAN など、他の典拠情報へのリンクを多数含む

・RDF として取得可能であり、SPARQL による問い合わせが可能

● 意義

Wikidata の最大の強みは、**異なる典拠情報を横断的に結びつけるハブとして機能しうる点**にある。実際、多くの人物項目には VIAF や ISNI、NDL Web Authority への対応関係が記述されており、これを利用することで、複数の典拠体系を一度に参照できる。

また、多言語ラベルが標準で管理されているため、国際的な研究プロジェクトや多言語資料を扱う場合には特に有用である。TEI/XML においても、`<idno type="Wikidata">` あるいは `@ref` 属性に Wikidata の URI を記述することで、外部知識基盤との接続点として活用できる。

さらに、生成 AI や検索拡張生成（RAG）の文脈では、Wikidata は構造化された背景知識の供給源として注目されており、TEI/XML で記述された人名情報と組み合わせることで、より豊かな文脈理解や関係推論を支える可能性がある。

一方で、Wikidata の利用には注意すべき点もある。最大の課題は、**情報の信頼性と一貫性が必ずしも保証されていない**ことである。Wikidata は誰でも編集可能な仕組みであるため、記述の精度や粒度にはばらつきがあり、専門的な人文学研究の観点からは不十分な場合も少なくない。

また、典拠管理という観点では、Wikidata は必ずしも「確定した正典」を提供するものではなく、複数の説や情報が並列的に記述されることもある。**そのため、研究においては、Wikidata の情報をそのまま正解として採用するのではなく、参照情報の一つとして位置づける**姿勢が求められる。

● TEI における位置づけ

以上を踏まえると、Wikidata は、NDL Web Authority や VIAF、ISNI、Getty ULAN のような**専門機関が管理する典拠情報を置き換えるものではない**。むしろ、TEI/XML による人名記述においては、人手で判断した人物同一性を `<listPerson>` と `@corresp` によって明示し、主要な典拠（NDL、VIAF、ISNI 等）を優先的に参照しつつ、Wikidata を **補助的かつ横断的な参照先**として併用する、という使い分けが現実的である。

このように、Wikidata は課題を抱えつつも、異なる知識体系を結びつける媒介として大きな可能性を持っており、TEI ガイドラインに基づく人名典拠の記述に組み込むことは今後大きな価値を持つことになると期待される。

2-4．TEI における典拠情報の位置づけ

TEI では、本文中の `<persName>` と、`<listPerson>` 内の `<person>` を、`@corresp` や `@ref` などの属性によって結びつけることで、**表記の揺れを保持したまま、人物同一性を明示的に記述し、外部典拠情報とも接続**できるという構造を実現できる。このような記述は、人手による判断を尊重しつつ、機械処理や検索、生成 AI による分析の基盤を整えるという二つの要請を同時に満たすものである。人名典拠情報の活用は、単なる補助情報ではなく、**構造化テキストを研究基盤として機能させるための中核的要素**であると言える。

3．複数ファイルから参照可能なリストの管理 ― XInclude の利用

Oxygen XML Editor を用いて TEI/XML ファイルを作成していると、人名・地名・書誌項目などを `<listPerson>` や `<listPlace>` といったリストとして整備していく場面がしばしば生じる。特に、書簡や日記、複数巻にわたる資料群など、**複数の TEI/XML ファイルで同じ人物や地名を繰り返し参照する場合**には、同一のリストを共有したいという要求が生じることが多い。

このような場合、各ファイルごとに同じリストを重複して記述すると、リストの更新や追加の手間が増えたり、同一人物に対して異なる `@xml:id` を付与してしまう危険があったり、人物同一性の管理が煩雑になってしまう、といった問題が生じやすい。

そこで有効なのが、**リスト部分を独立した TEI/XML ファイルとして切り出し、他のファイルから参照する方法**である。この方法を用いることで、リストの一元管理が可能となり、ID 管理や参照関係の整理が大幅に容易になる。

3-1．リスト専用ファイルの作成

まず、人名や地名などのリストだけを含む TEI/XML ファイルを新たに作成する。ここでは例として、ファイル名を linked_list.xml としておく。このファイルには、`<listPerson>` や

<listPlace> などを含め、@xml:id を付与した人物・地名情報のみを記述する。

この linked_list.xml 自体は、本文を持たない補助的な TEI ファイルとして位置づけられるが、TEI ガイドライン上は正当な使い方の一環と考えてよい。

3-2．XInclude を用いたリストの読み込み

次に、このリストを参照したい側の TEI/XML ファイルに、これを読み込めるようにするための設定を行う。ここでは、外部の XML ファイルを読み込むための仕組みである XInclude を利用する。具体的には、ルート要素（通常は <TEI>）に、以下のように XInclude 用の名前空間宣言を追加する。

```
xmlns:xi="http://www.w3.org/2001/XInclude"
```

この宣言により、当該文書内で xi: 接頭辞を用いた XInclude の記述が可能になる。

その上で、リストを読み込みたい適切な位置に、次のような要素を記述する。

```
<xi:include href="linked_list.xml"/>
```

これにより、linked_list.xml に含まれている内容が、その位置に展開されたものとして処理される。Oxygen XML Editor 上でも、XInclude を解決した状態で文書を扱うことができるため、外部ファイルに定義された @xml:id を、あたかも同一ファイル内に存在するかのように参照できるようになる。

3-3．XInclude を用いることの意義

この仕組みを用いると、概念的には次のような関係が成立する。

- 人名・地名などの定義は linked_list.xml に集約される
- 個々の本文ファイルでは、その定義を @corresp や @ref を用いて参照する
- 複数の TEI/XML ファイルが、同一の典拠リストを共有する

この方法により、人物や地名の追加・修正はリストファイル一箇所で行えばよくなり、結果としてデータ全体の一貫性が保たれる。また、@xml:id の重複や揺れを防ぐことができるため、後続の検索、XSLT 処理、Python や生成 AI による分析においても、大きな利点がある。

3-4．TEI における参照設計としての位置づけ

XInclude を用いたリストの外部化は、単なる作業上の工夫ではなく、TEI における参照設計

の一形態であると言える。

- 本文は本文として完結させる
- 同一性を管理すべき情報は集中管理する
- それらを明示的な参照関係で結びつける

という設計は、大規模化・長期化する研究プロジェクトにおいて特に重要である。人手で判断した人物同一性を、構造化された形で保存し、複数文書から共有するという点において、XInclude は TEI の考え方とよく整合している。

4．内部リストと外部典拠の役割分担

4-1．実務的観点から見た外部典拠利用の難しさ

人名や地名の同一性を管理するにあたって、VIAF や NDL Web Authority、Wikidata などの外部典拠を活用することは、多くの利点を持つ。しかし実務的には、外部典拠を起点として人物や地名を同定しようとすると、候補となる典拠を探索し、その妥当性を一つ一つ確認する作業に相当の時間を要してしまう場合が少なくない。特に、近代以前の人物や、特定の地域・資料群に限定された人物については、外部典拠側に十分な情報が存在しないことも多い。

そのため、実際の研究作業においては、**まず内部リストを作成し、人手で対応可能な範囲について研究者自身の判断によって人物・地名の同一性を確定させ、その後に必要に応じて外部典拠と接続していく**という方針が、効率面でも持続可能性の面でも有用である。この考え方は単なる作業上の妥協ではなく、TEI ガイドラインの設計思想ともよく整合している。

4-2．内部リストと外部典拠の役割の違い

内部リストとは、`<listPerson>` や `<listPlace>` に代表されるように、**当該研究プロジェクトの中で「誰を誰として扱うか」「どの名称を同一の実体としてまとめるか」を、明示的に記述したもの**である。ここには、本文読解や史料批判に基づく研究者の判断が直接反映される。略称、異表記、旧姓、通称といった資料固有の表現を柔軟に扱える点は、内部リストの大きな利点である。内部リストに付与される `@xml:id` は、プロジェクト内部で一意であれば十分であり、その意味づけと責任は研究者側にある。

これに対して、VIAF、NDL Web Authority、ISNI、Wikidata といった外部典拠は、プロジェクトの外部に存在する標準化された識別体系であり、異なる研究成果やデータセットを相互に接続するための参照点として機能する。外部典拠を参照することで、国際的・分野横断的なデータ連携が可能になり、検索、可視化、生成 AI を用いた分析などにおいても大きな利点が得られる。

ただし、外部典拠はあくまで参照先であり、研究上の判断そのものを代行するものではない。外部典拠に記載されている対応関係や属性情報は有用な手がかりを提供するが、それをどのように解釈し、どの範囲まで採用するかは、各研究プロジェクトの目的と責任に委ねられる。したがって、**外部典拠を先に固定的な基準として置くのではなく、内部リストで確定した研究判断を外部世界と接続するための補助的手段として用いる**という位置づけが適切である。

4-3．実装としての XInclude と段階的な外部接続

実践的には、まず内部リストにおいて人物や地名の同一性を定義し、本文中の `<persName>` や `<placeName>` から `@corresp` 等によって安定した参照関係を構築する。その上で、外部典拠が容易に特定できる場合や、将来的なデータ共有・連携が想定される場合に限って、`<idno>` や `@ref`、`@sameAs` を用いて外部典拠へのリンクを追加する、という段階的な運用が現実的である。TEI による記述は後付けの拡張に耐えるため、外部典拠へのリンクを将来的に追加する余地を残したまま、まずは研究に必要な最小限の構造化を進めることが可能である。

さらに、内部リストを実際に運用していくうえでは、前節で扱った **XInclude を用いたリストの外部化**が、効率的かつ有効な方法となる。人名や地名の内部リストを本文とは別の TEI/XML ファイルとして切り出し、複数の本文ファイルから参照するように設計すれば、人物・地名同一性に関する判断を一箇所に集約できる。これにより、リストの追加や修正を行う際の手間が大幅に軽減されるだけでなく、`@xml:id` の重複や不整合を防ぎやすくなる。

このように、内部リストを XInclude によって外部化して管理することは、単なる作業上の工夫ではなく、**研究判断を安定的に維持し、複数文書にわたって一貫した参照関係を保証するための設計上の選択**である。内部リストを起点として人物・地名の同一性を定義し、それを複数の TEI/XML 文書から共有しつつ、必要に応じて外部典拠と接続していくという運用は、実務的な効率と理論的な明確さの両立を可能にする。

以上のように、**内部リストを先行させ、外部典拠は目的と余力に応じて段階的に導入する**という方針は、実務的であると同時に、TEI ガイドラインが想定する「研究判断を明示的に記述し、共有可能な形で拡張していく」という考え方に即したものである。この役割分担を明確にすることによって、人手による判断、機械処理、国際的なデータ連携を無理なく両立させることができる。

5．まとめ

本章では、TEI ガイドラインを用いて人物名や地名といった固有表現を構造化し、同一性をどのように記述・参照・管理できるかを、演習を通じて具体的に確認した。本文中では `<persName>` や `<placeName>` によって表記揺れを含む言及を保持しつつ、`<listPerson>` や

<listPlace> によって人手による判断に基づく内部リストを整備し、@corresp などの参照属性によって両者を結びつけることで、研究判断を明示的に記録できることを示した。

また、NDL Web Authority や VIAF、ISNI、Wikidata などの外部典拠は、人物・地名を外部世界と接続するための重要な参照点となる一方、実務上は探索や検証に多くの時間を要するため、まず内部リストを先行して構築し、必要に応じて外部典拠と段階的に接続していく方針が現実的かつ持続可能であることを確認した。さらに、XInclude を用いて内部リストを外部ファイル化し、複数の TEI/XML 文書から共有する方法は、同一性判断を一元管理し、ID の一貫性を保つうえで極めて有効である。このように、本章で扱った手法は、単なるタグ付け作業ではなく、人文学研究における解釈判断を構造化し、機械処理や国際的データ連携、さらには生成 AI 時代の分析基盤へと接続するための基礎をなすものであり、次章で扱うより高度な構造設計や処理の議論への重要な足がかりとなる。

そして、TEI ガイドラインはとても有用なものだが、万能ではない。むしろ、TEI ガイドラインを使わない方が効率的にできる場合もある。あるいは、一通りデータを作成した後に、それを広く共有するために TEI ガイドラインに自動変換するという方法もある。クラウドソーシング翻刻サイトとして有名な「みんなで翻刻」でも、利用者側は意識せずに、TEI に準拠した出力機能を備えており、国立国会図書館の古典籍 OCR Lite も、TEI 出力機能を備えている。

このようにして作成した TEI 準拠のデータをどのようにして利用すればよいのか、ということについては、次章以降をご覧いただきたい。