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6章

構造化テキストの作成
――基礎演習 2――

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．より詳細な書誌情報の記述 ― 演習 1-4

学習の成果

この演習で習得すべき事項は以下の通りである。

- <teiHeader> の構造とメタデータを改良する
- <fileDesc> に含まれる以下の構成要素を理解する
  - ―出版と電子物の配布についての情報に関する <publicationStmt>
  - ―資料となる文書についてのメタデータを記録する <sourceDesc>
- <encodingDesc> を、ファイルの中で用いられているマークアップを記録するのに用いる
- <profileDesc> を、ファイルの書誌情報でない側面を記録するのに用いる
- <revisionDesc> で、ファイルへの大きな変更を記録する
- 名前を含む固有表現のタグ付けを深める

要点

この演習で学習するのは、まず、TEI/XML ファイルのヘッダをより深い意味を持たせる形で改良することである。そしてそれを通じて、そのマークアップと構造を理解することを目指す。これによって、<teiHeader> の様々な改良と、電子ファイルとその元資料に関する付加的なメタデータを記述する方法を体験することになる。さらに、名前を持つもののタグ付けにも取り組む。なお、TEI 準拠のデータ作成に際しては、ここで示す詳細なタグ付けは、分野や資料の性質によっては高い有用性を有するものの、必須ではないということに注意しておきたい。あくまでも、その資料を構造化する個人、あるいはコミュニティがそれを必要とし、かつ、そこにかかる時間的・人的コストを費やすことが可能である場合にのみ、実施すべき作業である。

①はじめに

まず、前回の演習で完成させたファイルを読み込んでみよう。もしそれが完成していない場合には、「soseki_letter_ex2_finished.xml」というファイルを読み込んで、他のファイルを保存しているフォルダに新しい名前で保存することで、少し近道をしてみよう。

② <publicationStmt> を改良する

ヘッダの中でも、<titleStmt> に関しては、演習 1-3 でいくつかの情報を記述した。しかし、その他の情報はまだかなり貧弱である。そこで、まずは <publicationStmt> を改良してみよう。このエレメントは、本来は様々な構造化された情報を記述できるが、今のところは一つの散文の段落しか持っていない。それをより詳細に書き換えてみよう。

1. <p> の開始タグ・終了タグを含む段落全体を削除しよう。

2. <publicationStmt> の内側に <publisher> エレメントを一つ追加しよう。ここで、Oxygen がどのように入力を補助するか、改めて確認されたい。なお、これは出版者（publisher）を記述するエレメントなので、今回の場合は自分の所属組織等を入力してもよいが、ここでは仮に「DH 大学 DH 研究科」としておこう。
3. <publisher> の次に、配布者を示す <distributor> エレメントを追加することもできる。ここでは仮に自分の名前を記述しておこう。
4. この後に、<authority> エレメントを追加しよう。これは、誰に出版・配布の権限があるかということを詳しく説明するためのものである。この場合には、自分の権限の下にある、ということで自分の名前を入れておこう。
5. 次に、<pubPlace> エレメントを記述し、さらにその中に <address> エレメントを一つ記述しよう。ここにはとりあえず仮の住所を書いてみよう。たとえば、<orgName> として「DH 大学 DH 研究科」、<street> の住所 (○○市◆◆ 4-11-8) を一つ、<settlement> (△△県) を一つ、<postcode> (999-9999) を一つ、<country> (日本) を一つ、入れてみよう。
6. <publicationStmt> の内側で、<pubPlace> エレメントに続けて <date> エレメントを追加し、そこに、たとえば「令和八年一月四日」という風に、本日の日付を漢数字で追加してみよう。<date> エレメントは @when 属性を用い、ISO による国際標準規格である ISO8601 に準拠して YYYY-MM-DD という日付の形式をとることができるので、<date when="2026-01-04"> のように対応する日付を記述しよう。
7. この後に、<idno> を使って ID 番号を追加しよう。これはカタログ番号のようなものや、この文書が属する URL となるべきものである。この場合、この手紙の自分の版にとって有益な ID 番号となると自分が考えるものをマークアップしてみよう。
8. 次に <availability> という記述を、この文書を配布する際に従おうと思うライセンスについての記述を含む <license> エレメントとともに追加してみよう。ここではクリエイティブ・コモンズライセンスの中からいずれかを選んで @target 属性に書いておくことを推奨する。（以下の例を参照。）
9. 現在の <publicationStmt> は、概ね以下のようになっているはずなので確認されたい。

```
<publicationStmt>
    <publisher>DH大学DH研究科</publisher>
    <distributor>自分の名前</distributor>
    <pubPlace>
        <address>
            <orgName>DH大学DH研究科</orgName>
            <street>〇〇市◆◆4-11-8</street>
            <settlement>△△県</settlement>
            <postCode>999-9999</postCode>
            <country>日本</country>
        </address>
    </pubPlace>
    <date when="2026-01-04">令和八年一月四日</date>
    <idno>N-S-0001</idno>
    <availability>
        <licence target="https://creativecommons.org/licenses/by-sa/4.0/">
            クリエイティブ・コモンズ　表示 - 継承 4.0 国際 (CC BY-SA 4.0)</licence>
    </availability>
</publicationStmt>
```

③ <sourceDesc> に詳細な書誌情報を記述する

現在の <sourceDesc> の記述も非常に貧弱である。ここには元になった資料の情報を書くことになっているが、今回扱っているのは書簡という 1 点物の資料であるため、本来は可能な限り丁寧に書くべきものである。この資料は東北大学附属図書館の Web サイトでデジタル画像として公開されているので、元資料である書簡自体の情報を記述することになる。この場合、手稿や稀覯本をはじめとする貴重な資料の書誌情報を詳細に記述するための <msDesc> を中心とするエレメント群が適していると思われるので、それを用いてみよう。なお、ここでは、<msDesc> の対象になる資料をまとめて「貴重資料」と呼ぶことにする。

1. <sourceDesc> のなかに現在ある段落全体を削除して、<msDesc> と置き換えよう。<msDesc> を入力すると、その内側に自動的に <msIdentifier> エレメントが追記される。<msDesc> は複製が広く流通する書籍とは異なり、その貴重資料を識別するための情報を必要とする。そのために <msIdentifier> が必須となっている。
   ① <msIdentifier> には、貴重資料を所蔵する組織、それが含まれるコレクション、そこでの識別番号等が含まれる。今回の漱石の書簡は、東北大学附属図書館の Web サイトで公開されているものであり、以下の Web 頁等で入力可能な情報を確認する。

```
https://touda.tohoku.ac.jp/collection/database/library/public/100300100003081
```

   ② 所蔵する組織については、<institution> エレメントで記述する。組織名だけでなく

住所等も記述できる。今回は、<orgName>エレメントに「東北大学附属図書館」、<address>の中に<settlement>を「宮城県仙台市」、<country>を「日本」と記述して<institution>以下に入れておく。

③　この書簡が含まれるコレクションは、「漱石文書」のようであり、さらにサブコレクションとして「2020年以降撮影」との記載がある。そこで、それぞれ<collection>エレメントに入れておき、サブコレクションに関しては@type属性で「sub」の値を与えておく。

④　識別子は<idno>に記述する。ここの組織での識別子は、まず、請求記号として「漱/29-1」があり、レコードIDとして「10030010003081」という値がある。さらに、デジタルコレクションとしてのURIと、それからIIIF Manifest URIも持っている。それらを@type属性で区別しつつ記述しておこう。

2. <msDesc>以下には、貴重資料に関する様々な情報を記述するためのエレメントが用意されている。内容について記述する<textLang>、資料の物理的な状況を記述するための<physDesc>などがあり、さらに、<physDesc>の中には、物理的構成要素を記述する<objectDesc>、筆致などを記述するための<handDesc>、綴じ方を記述するための<bindingDesc>、装幀について記述するための<decoDesc>、印章等について記述する<sealDesc>等、様々なものが記述できる。ここでは、<textLang>、<handDesc>、<bindingDesc>について。それぞれ以下のように気がついたことを記述してみよう。なお、<handNote scope="sole">における@scope属性のsoleという値は、それが一人の手で書かれたことを示している。その他、手書き資料についての詳細情報の記述の一部は、att.handFeatures属性クラスにまとめられているので参照されたい。

```
<sourceDesc>
    <msDesc>
        <msIdentifier>
            <institution>
                <orgName>東北大学附属図書館</orgName>
                <address><settlement>宮城県仙台市</settlement>
                        <country>Japan</country></address>
            </institution>
            <collection>漱石文庫</collection>
            <collection type="sub">2020年以降撮影</collection>
            <idno type="請求記号">漱/29-1</idno>
            <idno type="URI">
                https://touda.tohoku.ac.jp/collection/database/library/public/10030010003081
            </idno>
            <idno type="IIIF">
      https://touda.tohoku.ac.jp/collection/iiif/0/metadata/10030010003081/manifest.json
            </idno>
        </msIdentifier>
        <msContents>
            <textLang>口語に近い近代日本語の漢字仮名交じり文で書かれている。</textLang>
        </msContents>
        <physDesc>
            <objectDesc><supportDesc>
                    <extent><dimensions unit="mm">
                            <height>174</height>
                            <width>225</width>
                        </dimensions></extent>
                </supportDesc></objectDesc>
            <handDesc>
                <handNote scope="sole">やや崩した文字で書かれている。</handNote>
            </handDesc>
            <bindingDesc>
                <condition> 1枚紙の両面に書かれている。
                    この書簡が入っていた封筒も保存されている。 </condition>
            </bindingDesc>
        </physDesc>
    </msDesc>
</sourceDesc>
```

④ <fileDesc> の他の構成要素

ここまでは、<fileDesc> の下位で記述可能なエレメントをみてきた。ここまでほどに大規模なものではないが、他にもいくつか使えるものがあるのでみてみよう。

1. 終了タグ </titleStmt> の直後に、<editionStmt> を追記できる。これは、電子ファイル版に関して「第一版（First Edition）」<edition> のような説明的なフレーズを含む <edition> を記述するものである。ここでは第一版、と記述しておこう。

```
<editionStmt>
    <edition>第一版</edition>
</editionStmt>
```

2. 終了タグ </editionStmt> の直後に、<extent> エレメントを追記できる。これは、テキストの大きさの何らかの単位、たとえば、「260 字」や「2 ページ」といったことを書いて

おく。

3. 終了タグ </publicationStmt> の直後に、<notesStmt> を記述できる。これは一つ以上の <note> エレメントを内側に持つことができる。「構造化テキスト作成演習のために作成された」のようなものを含むことができる。

⑤ <encodingDesc> を追加する

<encodingDesc> エレメントは、その文書が実際にどのように構造化（エンコーディング）されたかを記述するためのものである。このエレメントは、この文書を他の人がコンピュータで処理・分析する際の拠り所となるものであり、可能な限り詳細に記述しておくことが重要である。

1. 終了タグ </fileDesc> の後に、<encodingDesc> を一つ追加しよう。
2. <encodingDesc> の内側に、<projectDesc> を追記しよう。その内側に <p> エレメントを入れて、「TEI ガイドラインによる構造化を学ぶための演習」と書いておこう。
3. 次に、<encodingDesc> の内側に、<editorialDecl> を追記しよう。このエレメントは、編集方針に関する情報を記述するために用意されており、その内容はさらに細分化できるように、いくつかのエレメントが用意されている。

   ここではまず、誤記に対する修正の方針を記述するために、<editorialDecl> の内側に <correction> エレメントを入力し、さらに <p> エレメントを追加して、そのテキストとして、「誤記と思われる文字列は <gi>sic</gi> を付与し、<gi>choice</gi> と <gi>corr</gi> で訂正を記した」と書いておこう。なお、TEI ガイドラインで XML 文中にエレメントを記述対象として書き込む場合には、半角の不等号記号を用いるのではなく、上述のように <gi>(generic identifier) エレメントを用いること。
4. 「正規化」に関する方針も記載しておくことが望ましい。ここでは <editorialDecl> の内側に <normalization> を入れてさらにそのなかに <p> エレメントを記入し、「漢字・ひらがな・カタカタ・合略仮名については Unicode14.0 の範囲で可能な限り元の表記に沿って記述した。変体仮名は使用していない。なお、合略仮名に関しては、<gi>choice</gi> と <gi>reg</gi> で対応するカタカナも示した。」というテキストを追記しよう。
5. このテキストデータでは、翻刻の段階から元資料には存在しない句読点を記述しているので、それについても説明が必要である。欧米圏であっても句読点が含まれない古いテキストが存在するため、これについてはエレメントが用意されている。<editorialDecl> の内側に <punctuation> と <p> エレメントを付与し「元資料には句点は存在しないが、翻刻者の判断により追加された。」と記述しておこう。あるいは、自分で句読点を追記・修正した場合は、そのことについても記述しておこう。
6. Oxygen が表示してくれる入力候補のエレメントリストを見て、<editoralDecl> と <encodingDesc> の内側で利用可能な他のオプションも確認してみよう。

7. あなたの <encodingDesc> は次のようになっているはずである。

```
<encodingDesc>
    <projectDesc>
        <p>TEIガイドラインによる符号化を学ぶための演習</p>
    </projectDesc>
    <editorialDecl>
        <correction>
            <p>誤記と思われる文字列は<gi>sic</gi>を付与し、
              <gi>choice</gi>と<gi>corr</gi>で訂正を記した</p>
        </correction>
        <normalization>
          <p>漢字・ひらがな・カタカタ・合略仮名については
            Unicode14.0の範囲で可能な限り元の表記に沿って記述した。
            変体仮名は使用していない。なお、合略仮名に関しては、
            <gi>choice</gi>と<gi>reg</gi>で対応するカタカナも示した。
          </p>
        </normalization>
      <punctuation><p>元資料には句点は存在しないが、
        翻刻者の判断により追加された。</p></punctuation>
    </editorialDecl>
</encodingDesc>
```

なお、<encodingDesc> は、文字の扱いをはじめとする様々な情報を記述することが可能であり、記述が手厚いほど、利用しやすくなる。自分にとっては当たり前のコンピュータ環境や資料の作り方であっても、他の人や、あるいは時代が変われば当たり前ではなくなってしまうことがある。そのような場合に、少しの手がかりだけでも残っていると有用なことがあるので、たとえ詳しいことは書けないとしても、データ作成した際に使ったソフトウェアを記述しておくだけでも役立つことがあるだろう。

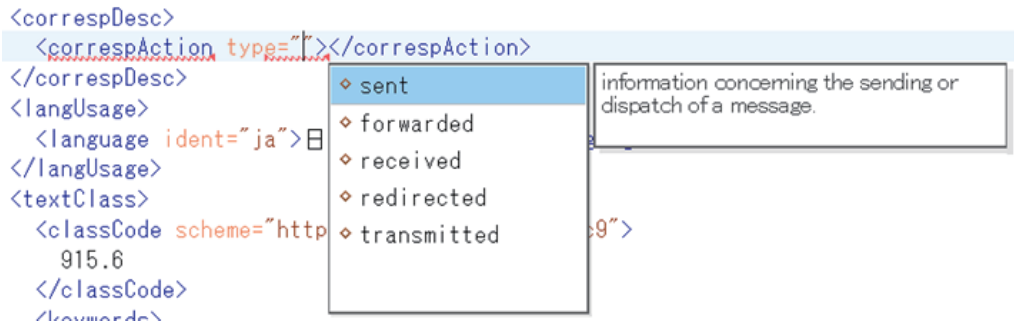
あるいは、データの機械可読性を高め、自動的に読み取りやすくするためには、<encodingDesc> 以下の記述をより精密に構造化するという方法もある。ただし、精密に決めようとすればするほど、例外的な事項が発生してしまい、構造的な記述が困難なケースが増えて、作業に多くの時間がかかってしまったり、結局のところ統一的なマークアップができなくなってしまう場合もある。そのようなことから、TEI ガイドラインでは、精密な記述方法を提供する一方で、上の例のように散文で自由に記述する方法も提供している。それでも敢えて、高度な自動処理を目指して精密な構造化に取り組むのであれば、一定のプロジェクトやコミュニティ等で議論してルールを共通化するとよいだろう。

⑥ <profileDesc> を追加する

<profileDesc> は、テキストに関する書誌情報以外の様々な情報を蓄積しておく場所である。ここにいくつかの情報を追加してみよう。

1. </encodingDesc> の後に <profileDesc> を追加しよう。
2. ここでは、書簡としての情報を記述してみよう。<correspDesc> は、Correspondence SIG が中心になって手紙の送受信を効率的に記述することを目的として 2015 年に TEI ガイドラインに追加されたエレメントである。送信と受信に分けてそれぞれについての情報を記述するだけでなく、転送等のいくつかのアクションが定義されているなど、手紙としての情報を取り出しやすいような記述手法となっている。

   ① まず、<correspDesc> を <profileDesc> の内側に入力してみよう。そして、<correspAction> エレメントを入力し、その後、@type 属性を入力しようとする。そうすると Oxygen は、その値として、sent, received, transmitted, redirected, forwarded の 5 つを提案してくる。ここではまず、送信に関する情報を記述するために sent を選んでおこう。



② これだけでは情報が不足しているため、妥当（valid）にはならず、エラーが出たままである。<correspDesc> エレメント群は、書簡のやりとりに関する情報を機械処理できるようにするために、少し厳密な記述ルールを定めていることから、これくらいの記述では、まだエラーとなってしまうのである。そこで、<correspAction> に必要なエレメントをさらに追加していこう。送信者として夏目漱石というテキストを含む <persName> と、日付情報を <date> エレメントを用いて以下のように書いてみよう。そして、送信地情報として <location> エレメントを追加し、その中に「ペナン」を含む <placeName> エレメントを追記しよう。そして、ペナンの港のあたりの座標情報を地図等で取得して <geo> エレメントで以下のように記述して、緯度・経度を空白区切りで記載しておく。

なお、この手紙の内容によれば、これが実際に発送されたのはこの後 10 月 1 日、コロンボに到着してからとのことだが、このときの漱石は汽船プロイセン号で欧州に向かっていた途上で、手紙自体は 9 月 27 日にペナン停泊中に書いたようである。この場合、発送をどの時点とみなすべきかについては議論の余地がありそうである。これについては、TEI ガイドラインでは「書いた日・場所」を別途記述できるので、それと送信日・場所を別々に

記述するか、それとも以下のようにペナンで書いた時点を送信日・場所と扱うか、ということを決める必要がある。これは編集方針として他の同様の書簡を構造化する際にも一貫して記述する必要があるので、それを踏まえた上でこの箇所をどう記述するべきか、検討されたい。

```
<correspAction type="sent">
  <persName>夏目漱石</persName>
  <date when="1900-09-27">九月二十七日</date>
  <location>
    <placeName>ペナン</placeName>
    <geo>5.418799642029674 100.345223003859</geo>
  </location>
</correspAction>
```

③　次に、受信者情報も入力しよう。今度は <correspAction> に続けてもう一つ <correspAction> を入力し、@type 属性を received としてみよう。受信地についてはこちらも <location> で、今回は住所がわかっているので <address> エレメントとその下位のエレメントを用いて書いておこう。また、座標情報は、現在は残っていない住宅であり、正確ではないが、大体このあたりということで <geo> エレメントで記載してみている。受信した日付は残念ながら不明であった。これらを一通り記述すると以下のようになる。

```
<correspAction type="received">
  <persName>夏目鏡子</persName>
  <location>
    <geo>35.70348307829794 139.73306153078642</geo>
    <address>
      <country>日本</country>
      <region>東京</region>
      <settlement>牛込區</settlement>
      <street>矢来町三番地中ノ丸丙八十五号</street>
    </address>
  </location>
</correspAction>
```

3. 次に、<profileDesc> のなかに、<handNotes> を追記しよう。この内側には <handNote> エレメントを記入し、そのテキストとして「漱石筆」と記述しておこう。
4. 次に、<langUsage> を、<profileDesc> の内側に、日本語の言語コードである「ja」を値として持つ @ident 属性を持つ <language> 日本語（明治時代）とともに付け加えよう。
5. <correspDesc> に続けて、<textClass> エレメントを追加しよう。そして、<classCode> エレメントをその内側に記述し、そのテキストとして「915.6」を、「http://jla.or.jp/data/

ndc9」という値を持つ @scheme 属性を追記しておこう。これは日本十進分類法での「日本文学・書簡・明治以後」を表すコードである。そして、これに続けて、十進分類法での915.6 に割り当てられているキーワードも <keyword> タグで記載しておこう。これは、十進分類法で探したい人や、それに対応したシステムにこのデータを組み込む際に役立つ。(なお、これはあくまでも一例であり、このコードの利用を推奨しているわけではない。)

6. あなたの <profileDesc> は、今、以下のようになっているはずである。

```
        </correspDesc>
        <langUsage>
            <language ident="ja">日本語 (明治時代) </language>
        </langUsage>
        <textClass>
            <classCode scheme="http://jla.or.jp/data/ndc9">
                915.6
            </classCode>
            <keywords>
                <term>日本文学</term>
                <term>書簡</term>
                <term>近代：明治以後</term>
            </keywords>
        </textClass>
    </profileDesc>
```

⑦ <revisionDesc> を追記する

<revisionDesc> はファイルへの大きな変更の情報を記録する方法を提供している。

1. 終了タグ </profileDesc> の直後に <revisionDesc> エレメントを追記しよう。
2. この電子データには少なくとも三回の revision の機会があったため、ここではそれを記述するために二つの <change> エレメントをこの内側に追加しよう。最初の一つに関しては、@when 属性を今日の日付とともに追記しよう。そして、<change> の内側で、あなたの名前を含む <persName> を追記しよう。そして「ヘッダの改良」というテキストを記載しよう。
3. 二つ目の <change> では、「2022-01-31」という値を持つ @when 属性を、「永崎研宣」の <persName> とともに追記し、「デジタル翻刻」と書いておこう。<title> として「漱石筆鏡子宛て書簡」をマークしておくのもよいだろう。
4. ここでは、一番新しい <change> が最初に来るのが一般的である。
5. <revisionDesc> は、今、次のようになっているはずである。

⑧自分の作品を保存する

自分の作品を保存しよう。

・現在の作品は形式が整っているか？幸せの緑の四角と怒りの赤い四角のどちらが出ているか？
・自分の作品を自動的に整形してインデントしたか？
・「ファイル」メニューから「保存」を選ぶか保存アイコンをクリックするか、あるいは、「ファイル」を選んでから「別名で保存」のメニューを選んで、「soseki_letter_ex3.xml」の名前を用いるか、あるいは好きな名前を使って保存しておこう。

⑨セルフチェック

以下の質問に答えることで、今回の演習の中核的な原理を理解しているかどうか確認してみよう。

・どんな種類のメタデータを `<titleStmt>` の中に蓄積することができるか？
・`<publicationStmt>` は何のために使われるか？それは何を含むことができるか？
・ファイルに関する元資料の詳細情報をどのようにして提供するか？
・書簡に固有の構造はどのように捉えられているか？
・`<encodingDesc>` は何のためのものか？
・`<change>` エレメントは `<revisionDesc>` の中ではどんな順序で並べられるべきか？

⑩もっと深めたい人へ

・今回使った新しいエレメントのそれぞれに関して TEI P5 ガイドラインのエレメントの説明のページを見てみよう。
・ガイドライン第 2 章、teiHeader に関する章を読んでみよう[1)]。
・

注

1　https://tei-c.org/release/doc/tei-p5-doc/ja/html/ref-att.handFeatures.html