# 構造化テキストデータの活用における現状と課題

## —TEI に準拠した『浄土真宗聖典全書』全文検索システムの開発を通じて—

永崎研宣（一般財団法人人文情報学研究所）
中村覚（東京大学史料編纂所）
田中真　西河雅人　林龍樹　井上慶淳（浄土真宗本願寺派総合研究所）
下田正弘（東京大学大学院人文社会系研究科）

**概要**：『浄土真宗聖典全書』は、Procedural markup を志向した独自タグセットによる構造化テキストデータとして紙媒体で印刷・刊行された。本稿は、そこで用いられたデータを国際標準である TEI ガイドラインに準拠した Descriptive markup へと変換して利便性の高い全文検索システムを開発した際の、テキストデータ変換における現状と課題について議論する。変換は Python を用いたプログラムを開発することで機械的に行われたが、この種の変換ではある程度までは機械処理での変換が可能であるものの、それを踏まえた上で、記述の対象となる構造の選択はあくまでもシステムを作成する人や組織の側に委ねられていることを確認した。
**キーワード**：仏教学, TEI ガイドライン, テキスト構造化, 全文検索

# Current Status and Issues in Utilization of Structured Text Data:

## Development of a Full-Text Search System for TEI-Compliant Buddhist Scriptures

Kiyonori Nagasaki (International Institute for Digital Humanities)
Satoru Nakamura (The University of Tokyo)
Makoto Tanaka, Masato Nishikawa, Ryuju Hayashi, Keijun Inoue
(Jodo Shinshu Honganji-ha Research Institute)
Masahiro Shimoda (The University of Tokyo)

***Abstract***: Jodo Shinshu Seiten Zensho (A Collection of Jodo Shinshu Writings) has been printed and published in paper form as structured text data with an original tag set oriented toward Procedural markup. This paper discusses the current status and issues in converting text data used in the books to descriptive markup that conforms to the TEI guidelines, an international standard, to develop a highly convenient full-text search system. The conversion was done mechanically by developing a program using Python. Although this type of conversion can be done by machine processing to some extent, it was confirmed that, based on this, the choice of the target structure of the description is solely left to the person or organization creating the system.
***Keywords***: Buddhist studies, TEI Guidelines, Text encoding, Full-text search

## １. まえがき

本発表で取り扱うのは、発表者らが取り組んできた浄土真宗聖典全書の全文検索システムの開発における技術的課題とその解決策、そして今後の課題である。この取組みでは、紙媒体で印刷するための版面を前提とした精緻なマークアップを付したテキストデータがすでに整備されており、これを、テキストの内的構造を前提とした形式に変換しつつ Web 上で全文検索可能なものとすることを目指した。そこで、このプロセスにおいて生じた技術的課題を挙げた上で、その解決策、及び残された課題について論じる。

## ２. 対象となる資料とデータ

本発表における対象資料である『浄土真宗聖典全書』は、浄土真宗本願寺派総合研究所が 14 年（2005-2018）に亘り編纂した浄土真宗聖典の全書（全 6 巻）である。浄土真宗の正依の経典たる浄土三部経や異訳大経・小経と七高僧の論釈、宗祖親鸞の撰述を収録しており、さらにその教えを伝承し意義を明らかにした覚如・蓮如等の撰述、教義の理解に重用される関連典籍や史資料を網羅した浄土真宗聖典の集大成である。これを効果的に構造化し全文検索可能とすることが本研究の目標であり、それに際して用意されたテキストデータについて以下に述べる。

### 2.1 元となったテキストデータの概要

『浄土真宗聖典全書』は、当初、紙面の組版を前提とした独自のタグセットを策定し、それを付与する形でテキストデータが作成されていた。(以下、このデータを「元データ」と呼ぶ)このタグセットは、本文では用いられない全角の記号とアルファベットを基調とし、紙面の詳細なレイアウトの指定を可能としている。タグの付与の仕方としては、本文の対象文字の前後に全角の記号やアルファベットを付記し、その中に右仮名・左仮名・返点・句読点、そして校異番号などを付加している。原典の改頁や改行、インデント、偈文などを示すタグもあり、史料の形態についても一定程度記述可能である。校異情報に関しては、別ファイルとして作成し、本文中の校異番号と紐付ける形で内容を記述している。校異として用いられる各典拠資料については◎・甲・乙など、印刷版で用いられる記号をそのまま利用して表記し、印刷版にて対応づけを視認できるようにしている。また、文字コードとしては Unicode（UTF-8）を用いており、字体は典拠資料の字体に依拠しつつ、主に旧字体の通行体（基本字）を採用し、異体字は用いていない。

### 2.2 元データにおける課題

元データのタグセット及びそれに基づいて作成されたテキストデータは書籍としての刊行を前提として設計されたものであり、章タイトルや本文左右の注や脚注などを版面上に適切に配置することに際しては十全なデータである。しかしながら、Web で公開し検索を効果的に行えるようにするためには、書式としては同じ形になっている各種の注に含まれる内容を機械的に峻別できることが望ましく、また、それも含めた全体的なテキストの内容的な構造が適切に記述されていることが望ましい。

この相違は、テキスト構造化の議論においては、Procedural markup（手続き的マークアップ）と Descriptive markup（記述的マークアップ）として区別されるものである。前者は主に印刷における見た目の再現を目指してマークアップする方向性であり、後者は内容の記述を目指すものである[1]。元データはいわば Procedural markup にあたるものであり、内容を踏まえた検索や分析を行うためには、Descriptive markup に沿って構造化を行うことが有用ということになる。

Descriptive markup を目指すにあたっては、どのような内容に着目し、どのようにしてマークアップを行うかを決定する必要がある。すなわち、扱うデータ群において着目する要素やそのマークアップの内容的な深さをそろえておかなければ、構造化データと言っても効果的な処理や分析が困難になってしまう。この点は特に複数人でデータ構造化を行う場合には注意しなければならない点である。このような事柄を検討する上で参考になるのは、TEI 協会図書館分科会が公表している Best Practices for TEI in Libraries[2]である。ここでは、5 段階のマークアップのレベルとそれぞれに対応する推奨エレメントを提示し、構造化の深さを策定し共有するための手がかりを提供している。レベル 1 では OCR をかけたテキストをほぼそのまま提示するマークアップとなっており、レベル 2 ではごく最小限のマークアップで章タイトル等がわかるようになっている。ここまではいわば Procedural Markup と言えなくもないようなものだが、レベル 3 では、ごく簡素ではあるものの、テキストの内容的な構造を踏まえたマークアップが提示されている。たとえば、散文であれば段落毎に<p>エレメントを付与し、韻律詩であれば<l>エレメントを付与することになる。これは OCR をかけただけでは判断がつかないことであり、人手でのマークアップを前提とすることになる。そして、レベル 4 では、テキストへの追記や修正、登場する人名や戯曲の各台詞の話者など、一般的に広く利用され得る基礎的なマークアップを行うとする。最後のレベル 5 は、研究者各自による独自の深いマークアップとされており、特に個別の内容は提示されていない。このレベルであればもはや詳細を示す必要がないということだろう。

このようなことを踏まえると、検索の利便性を高めるためにマークアップを行うとしたら、ここで言うレベル 3 もしくは 4 を目指すということになる。

一方、別の観点からの課題として、元データは、XML や JSON 等の既存の記法とも異なる独自のものとなっていたため、付与したタグの構造の検証を十分に行うことができず、タグのミスの発見があまり容易ではないこと、新たな独自記法のため、新たな課題が生じるたびにタグやその記法が変化してしまい、記法の一貫性の確保が難しいこと等があった。これらは、XML 等の標準化された記法であれば機械的な検証が可能であることも多く、既存のある程度

充実した構造化手法を用いれば記法が変化することもかなり少なくなるため、このような課題を解消するには、既存の何らかの記法に変換することが一つの有力な解決策である。

## 3．記述方式の変換

上述の課題を解決するに際し、元データの性質を勘案した結果、筆者らは、元データを TEI (Text Encoding Initiative)ガイドライン（以下、TEI-G）[3]に準拠した形式に変換することとした。

この理由としては、まず、データの構造に関して、テキスト本文の任意の箇所に注や脚注記号等が部分的に付与されることになるため、一つの階層構造に収斂可能なものではなく、いわゆるインラインタグと呼ばれる、一連のテキストデータの任意の場所にタグを付与できる仕組みを持っている必要があり、さらに、テキストの内的構造だけでなく、それと異なる改行や改頁等の版面の構造も記述する必要がある。したがって、JSON や XML 等による素朴なツリー構造のデータ形式では対応が困難である。一方で、ツリー構造の弱点を克服し得るグラフ構造とする場合には、テキストデータの任意の箇所をノードとしなければならないため、シンプルなグラフ構造に落とし込むことは現在の処理系ではなかなか容易ではない。このようなことから、折衷的な方法として、XML においてインラインタグを可能としつつ、主となるツリー構造以外は空白タグ（Empty element）を用いて表現する方法や、さらに、離れたタグ同士の関係を XML 属性における ID 参照等によって記述する手法が用いられることがある。TEI-G はこの手法を積極的に活用しており、すでに広い実績を有していることから、記述手法として今回の目的にふさわしいものであると考えられた。

データの内容的な構造を記述するための構造化手法としての Descriptive markup について、TEI-G はその始まりである 1987 年より取組みを開始し、今なお改良を続けてきていること、そして、一方で、Web 上での表示と紙媒体での版面を対応づけるという要請もあり、Procedural markup の共存が必要となったが、その点についても TEI-G は長い蓄積と豊富な手法を提供していることから、今回の件には適切であると考えられた。

前章で挙げた独自タグセットにおける課題に関しては、TEI-G の場合には、国際標準規格として 20 年以上に渡り利用され MS-Office をはじめ様々なアプリケーションのデータフォーマットに採用されることにより多数のユーザを抱える XML に準拠していることから、様々な課題への対処方法がすでに記述言語のレベルで提供済みである。たとえば、XML タグの整合性の確認を自動的に行うツールはフリーソフトとして提供されており、それを組み込んで編集作業中のデータを随時自動検証するエディタも複数存在する。プログラミング言語においても、一般によく用いられるものであれば、XML をパースして効率的に処理するためのライブラリやモジュールが提供されている。

そして、XML のサブセットの一つである TEI-G としては、SGML 準拠のものに始まる 30 年以上の歴史的経緯を踏まえ、現在ではタグセットの名前空間をはじめとする構造を規定した XML スキーマを配布していることから、TEI-G が定義するタグの整合性の検証も自動的に実行可能である。また、人文学に関わる様々な種類の文書に対応したタグセットとして長く蓄積されてきた知見に基づいて多くのタグとその用法が整備されており、特に、今回扱う多様な注や校異情報を含む文献学的な構造を必要とするデータに関しては、元資料の多様さに十分対応できるようなエレメントと属性、そしてその階層構造の制約が定義されている。それでも不足する場合には一時的に付与するための汎用タグが用意されていることから、タグの記法が変化する可能性を最小限に抑えることができる。そして、TEI-G に準拠した表示・検索・分析等を行うためのツールやシステムは、XML のライブラリやモジュールを利用できるため比較的開発が容易であり、世界中で様々なものが開発公開されている。欧米言語向けに開発されたものであっても UTF-8 に対応しているものであれば日本語資料でもほぼ利用可能なため、利用可能なツールの種類は幅広い。そして、発表者らもまた、日本語・東アジア言語資料向けの TEI-G テキスト向けのものを開発・公開している。

このようなことから、今回の全文検索システムの開発にあたっては、元データを TEI-G に変換することで実現することとした。

### 3.1 記述方式変換における課題

元データの独自マークアップは、その内容を可能な限り TEI-G に反映させることを目標とし

て自動変換を行うことが望ましい。そこで、元データの独自タグセットとこれに対応する TEI-G のタグとのすりあわせがまずは必要であった。その結果、すべての独自タグを TEI-G に置き換え可能であることは確認できた。仏典への TEI-G に準拠した構造化に関しては、すでに様々な種類の仏典を含む大正新脩大藏経の TEI-G 準拠に向けた作業[4]が先行しており、それを踏まえることで TEI-G におけるタグや構造の選択については大きな問題なく対応可能であった。

とりわけ、左右注や送り仮名に関する情報については、元データでは精密に記述されており、この種の形式については適切な変換が可能である。

一方、元データはあくまでも印刷版面の表現を主な目的としたものであり、各文献の解題において文章で説明したり、一定の形式で提示した後は読者の判断に委ねることになる。そのような部分については、内的構造を踏まえた TEI-G 準拠への自動変換が困難なこともある。その場合、当面は、元データの記述によって表現可能な範囲でとどめておき、今後の課題としておくこととした。一例を挙げると、手紙（消息）のような典拠が一つだけの文献においては編集者による誤字や欠字の修正が行われる一方で、複数の典拠を持つ文献の場合には、いずれかの典拠に基づく批判的な校訂が行われる。TEI-G では前者は<choice>タグを用いて以下のように記述し、

```
<choice>
  <sic>明らかな誤りを含む原文ママ</sic>
  <corr>原文を修正したテキスト</corr>
</choice>
```

後者は<app>タグを用いて以下のように記述するルールとなっている。

```
<app>
  <lem wit="#原本">本文</lem>
  <rdg wit="#甲">異読 1</rdg>
  <rdg wit="#乙">異読 2</rdg>
</app>
```

紙媒体に印刷した場合、これらは同じ形式で記述可能であることから、元データでは、個々の文献の解題の箇所でその文献における脚注に位置づけについて文章で説明することで用が足りており、脚注データの形式としては両者が同じ形式となっている。紙媒体版としてはそれで十分だが、デジタルデータとしての活用可能性という観点では、それらを区別することで、より利便性と信頼性の高いデジタルデータの構築を目指すことが望ましい。

### 3. 2 変換の実際

今回の変換では、Python3 の etree 及び lxml モジュールを全面的に利用した。これを用いて、元データの各タグに対応する TEI-G のエレメントや属性を生成した上で、DOM ツリーを構築していった。そして、最終的に、TEI-G の XML スキーマを用いて jing で validation を行った。

元データでは、例えば図 1 のように、ふりがなや送り仮名、漢文記号等、本文に対する様々な付加情報がタグを用いて付与されており、これが紙媒体では図 3 のように印刷されていた。この種のタグは非常に多く付けられており、通常のインラインタグ形式ではタグのオーバーラップが生じる可能性が考えられた。そのため、それぞれの形式に応じた@type 属性を付与した<note>タグを用いてその種の付加情報を対象となる本文文字の始点に記述し、対象文字（列）の終点に ID 付きの<anchor/>を置いてそれを<note>タグから ID 参照する形式とした。（図 2）これを全文検索システム上で表示すると図 4 のようになる。

成〔じやう／〕佛〔ぶち／より〕

図 1: 元データの記述の例

```
<note xml:id="nt_79" targetEnd="#nt_79e" place="right">
    じやう
</note>成<anchor type="noteEnd" xml:id="nt_79e"/>
<note xml:id="nt_80" targetEnd="#nt_80e" place="right">
    ぶち
    <metamark function="okuri">より</metamark>
</note>佛<anchor type="noteEnd" xml:id="nt_80e"/>
```

図 2: 図 1 に対応する TEI-G 準拠データ

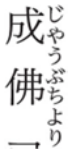


図 3: 図 1 に対応する紙媒体の版面


図 4: 図 2 に対応する Web 表示画面

なお、この<note>の用法は、「廣瀬本万葉集」において用いたもの[5]を踏襲しており、日本語古典籍 TEI 本文データ作成要領にも一例としてあげられている[6]。

校異情報データの記述に関しては、TEI-G では、(1) Location-referenced method、(2) Double-end-point-attached method、(3) parallel segmentation method の三種の記述手法が提示されている[7]。(1)は紙媒体の書籍等で脚注記号が本文に記載されているのと同様に、校異情報と何らかの形で関連のある箇所を示し、そこに関する校異情報を記載する手法である。この手法は、表示は適切にできるものの、本文と異文の対応情報を取得しづらいためそれらの関係を機械的に処理するといった用途にはあまり適していない。(2)は、校異情報における異文（異読）のある箇所の始点と終点にアンカータグ等を設置し、その 2 点を参照する形で異文を記述する手法である。この手法では、異文と本文との対応関係を機械的に取得・処理することができ、切り替えたり、統計処理を行ったりすることが比較的容易である。さらに、タグのオーバーラップを避けられるというメリットがある。その一方で、XML のツリー構造を活用したものではなく、ID の参照関係を用いて処理する必要があるために、論理的には処理可能であるものの、処理するためのプログラミングにはやや複雑になる。また、(1)と(2)では、異文情報は、本文の外に記述することができる。(3)は、校異情報を記載する本文にインラインタグで異文も含めて記載するものであり、タグがオーバーラップしやすいという課題があるものの、XML のツリー構造に基づいて記述するため、表示を含む処理に際しては簡素なプログラムで対応可能である。

このような特徴と元データの状態を踏まえ、校異情報の記述方法は、(2) Double-Endpoint Attachment method を前提とした (1) Location-referenced method を採用した。すなわち、元データでは校異情報は本文とは異なるファイルに頁番号・脚注番号とともに記述され、本文中の脚注番号とリンクされている。そのため、校異情報は<back>に<listApp>を配置し、そこに<app>を列挙しつつ、本文に配置したアンカータグを ID 参照する形とした。元データの本文自体は精度が高く、校異情報には本文のテキストも入っているため、理論的には機械的に Double-Endpoint Attachment method での記述が可能なはずだが、タグを超える等のいくつかの処理が必要になる可能性があるため、当面は Double-Endpoint Attachment method の始点と終点のタグの双方を始点箇所（ここに元データにおける脚注記号が入っている）に配置する形で処理し、他の課題が完了した後に、改めて終点の位置を処理することとした。したがって、現時点での構造としては Location-referenced method ということになった。

このようにして、当該全集に含まれる文献の単位で XML 文書を作成し、全体では 221 件となった。先述した<choice>と<app>の区別に関しては、文献単位で異なっているものがほとんどであるため、元データに手を加えて区別できるようにすることで大部分については機械的に対応可能であった。

## 4. 表示と検索

全文検索システムとしては、デジタル源氏物語や延喜式等のために作成した Web ビューワ[8][9]で縦書きや校異情報表示をすでに中村が実現しており、それをほぼそのまま利用可能であった。ただし、<choice>と<app>の表示に関しては、ともに脚注として表示し、その内容は人の目で区別する仕組みとした。データ表示の全体的な仕組みとしては、JavaScript を用いて XML をいったんエレメントごとに JSON に変換し、それを Vue.js のコンポーネントに組み込んで表示している。この手法は、Raffaele Vigilante が React を用いて実践しているもの[10]を参考にしている。（図 5）


図 5: 本文及び校異情報表示の例

検索に際しては、LAMP 環境という制約から、サーバサイドでは Laravel を用い、PHP に MySQL を組み合わせて作成している。データに旧字体を用いていることから、検索の便のために新字体での検索も可能とする必要がある。これについては、検索システム側で対応しており、異体字検索に関しては独自の対応表を作成して用いている。また、左右の仮名等、タグ付けした情報を用いた絞り込み検索や、Ngram Viewer 検索等も提供している。

## 5. 今後の課題

本発表では、版面の構造を前提としたテキストの構造を、テキストの内容に基づいた構造に変換したことに焦点をあてた。取組みとしての当初の目標である全文検索システムは作成できたものの、見た目に基づくマークアップを完全に内容構造へと置き換えられたわけではなく、前述のもの以外にも課題として残るものがある。人が読んで理解することを前提とした紙媒体の資料は、内容の構造に関する区別を読者に任せることが相当程度可能であり、これは Web においても人が閲覧することを前提とした場合には同様に考えることも可能である。一方で、TEI-G の側では必ずしも深い内容に基づく構造化を要求しているわけではなく、深い構造化も浅い構造化も同様に選択可能であり、さらに、見た目に関するタグも TEI-G では様々に用意されている。どのような構造化を行うかの判断はあくまでもデータ作成者に委ねられている。本研究では、全文検索システムとしての実用性を確保しつつ、自動変換と手動作業のバランスに配慮しながら、今後も、より深い構造化に基づく利便性の向上を目指していきたい。

## 謝辞

本研究の一部は、JSPS 科研費 JP19H00516 の助成を受けたものです。

## 参考文献

[1] Renear, Allen H. Text Encoding. A Companion to Digital Humanities, 218‐39. Blackwell Publishing, 2004 年.
[2] Best Practices for TEI in Libraries, https://candra.dhii.jp/nagasaki/TEI/bptl-driver.html (2022-10-30 accessed).
[3] The TEI-Guidelines, https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html (2022-10-30 accessed).
[4] 渡邊要一郎.「大正新脩大藏経 TEI 化に関する概略」. 一般財団法人人文情報学研究所(監修)『人文学のためのテキストデータ構築入門』, 第 4 部 事例編, 316-48. 文学通信, 2022 年.
[5] 永崎研宣他. 万葉集伝本研究のためのデジタル基盤構築. 情処研報, 2021-CH-125, 2021 年.
[6] 日本語古典籍 TEI 本文データ作成要領, https://github.com/TEI-EAJ/jpn_classical/blob/master/jpn_classical_guideline.md (2022-10-30 accessed).
[7] 12.2 Linking the Apparatus to the Text, 12 Critical Apparatus, The TEI-Guidelines. https://www.tei-c.org/release/doc/tei-p5-doc/ja/html/TC.html#TCAPLK (2022-10-30 accessed).
[8] 中村覚他. デジタル源氏物語（AI 画像検索版）. 情処研報, 2022-CH-128 2022 年.
[9] 小風尚樹他. 相互運用性を高めた日本歴史資料データ実装. じんもんこん 2021 論文集, 2021:294‐301. 2021 年.
[10] simple-ceteicean-react, https://github.com/raffazizzi/simple-ceteicean-react (2022-10-30 accessed).