構造化記述されたテクストの基盤整備に向けて：延喜式の TEI マークアップを事例に

後藤　真（国立歴史民俗博物館）　小風尚樹（東京大学大学院，キングスカレッジ・ロンドン）
橋本雄太（国立歴史民俗博物館）　小風綾乃（お茶の水女子大学大学院）
永崎研宣（（財）人文情報学研究所）

本研究は，日本古代の史料である『延喜式』に対して TEI マークアップを施した際の，マニュアル等の記述の状況について述べたものである．TEI は，広く人文研究のためのテクストを作る国際標準として重要であるものの，そのルールが複雑であり，専門家以外には記述が困難であるという状況がある．また，このようなデータを作成した際に「どのような意図でデータを作ったのか」を記録することで，研究そのものをトレースすることができるようになるとともに，データの長期保存にとっても有益であると考えられる．

An approach to establish text infrastructure with structured description

Makoto Goto (National Museum of Japanese History) Naoki Kokaze (University of Tokyo)
Yuta Hashimoto (National Museum of Japanese History) Ayano Kokaze (Ochanomizu University)
Kiyonori Nagasaki (International Institute for DigitalHumanities)

This paper describes a situation of a compilation of a manual for encoding an ancient Japanese historical source, Engi-Shiki, in compliance with the TEI (Text Encoding Initiative) P5 guidelines. The reasons why we construct the manual are that (1) the way of the markup can be too complicated for people who are not familiar with the TEI guidelines, (2) the process of the research can be easily traced by recording the intention of producing the data, and (3) it should contribute long-term preservation.

１．はじめに

本報告では，日本の諸研究機関における構造化テクストデータ蓄積と長期的な維持のための手法について述べる．筆者らは，これまでに日本古代の歴史資料である『延喜式』の TEI マークアップ化を行ってきた．そのプロジェクトの成果および課題をもとに，基盤構築のための手法等について提案を行う．

２．本研究の位置

『延喜式』は，藤原忠平・時平によって作成された，律令を補完し，施行細則を定めた法律である．この施行細則を定めた「式」は，『弘仁式』・『貞観式』に続き，三つ目の式としてこの『延喜式』が作成され，927 年に完成したものである．

この『延喜式』は，律令制度が崩壊しつつある時期のものであるため，いくつかの議論はあるものの，最初から最後まで全て一貫して残された式はこの『延喜式』以外には存在しないという点と，貞観式・弘仁式の一部を取り込んだ上で作成されているという点から，日本古代の状況を知る上では，資料価値の大変に高いものである．

『延喜式』は，古くは虎尾俊哉の研究から始まり[1]，多くの日本古代史研究者や，様々な研究者によって用いられている．人文情報学に関わる部分では，例えば，『延喜式』「神名帳」の神社の情報をベースとした歴史地名のデータが人間文化研究機構より公開されている．また，これ以外にも，『延喜式』の情報は金属加工・食品などの情報を含むなど，日本古代の様々な構造を知ることができる．

しかし，今の段階では，電子データとして閲覧可能なものは，国立国会図書館のデジタルライブラリーにあるに止まり，画像データでの閲覧が限界であった．また，関連する先行プロジェクトとして，カリフォルニア大学バークレー校の提供する Japanese Historical Text Initiative では，『延喜式』の 1～10 巻までの本文と英訳，対応する画像をウェブ上で閲覧できるようになっている点では非常に価値が高い[2]．ただ，残念ながら全てのテクストがあるわけではない．

そこで，人間文化研究機構が行う「異分野融合による「総合書物学」の構築」のプロジェクトおよび「総合資料学の創成」プロジェクトのもと，国立歴史民俗博物館においてテクストのデジタルデータの構築を行うこととなった．そのデジタル化に合わせて，著者らは TEI として作成することを提案した．『延喜式』全 50 巻を対象にマークアップを行い，利用者の研究関心に応じたデータ提供を目指すものである．

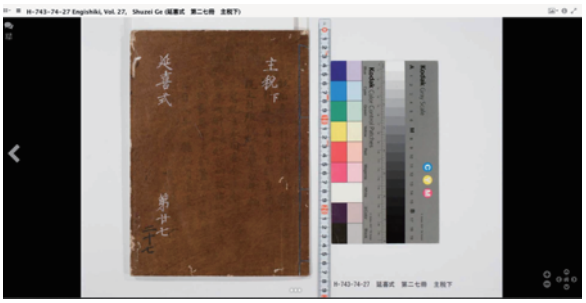


図 1　『延喜式』の例（IIIF の実験画像）

その結果，著者のうち小風は，トランザクショノグラフィの手法を用いた[3]，『延喜式』における物品のありようや，財政規模の分析の可能性を提案した[4]．さらにその課題の中から，TEI における度量衡の単位を表現するための適切なエレメントが存在しないことを指摘し，それを提案するに至った[5]．上記のような状況の中で『延喜式』研究における TEI の有用性は，一定程度見込まれるであろうことは判明した．しかし，『延喜式』をより広範に研究として用いるためには，より基盤となりうるようなマークアップをほどこしたデータを作ることが必要である．

一方，これまで，日本における TEI マークアップ手法の検討としては，高橋洋成などの仕事がある[6]．これらの研究は，ある個別の研究目的に即したものであり，それ自体は十分に価値と意義を認めうるものであるが，すぐに基盤となるマークアップとなるものではない．著者の一人でもある永崎は大正新脩大蔵経データベースに TEI を適用する検討を行なっている[7]．これは基盤的なテクストデータへの TEI 適用という数少ない事例ではあるが，その手法自体が必ずしも共有されていない部分がある．

3．基盤となるテクストデータ構築のためのマニュアル

そこで，本研究では，基盤となりうるデータ構築の手法そのものの共有化を測ることを目指すこととした．具体的には，マークアップ作業の記録をもとに，それをマニュアルとして整備し，共有することを目指す[8]．

TEI それ自体が，テクストをどのように認識するかを可視化し，共有化するための手法である．しかし，実際には TEI は極めて複雑であり，簡単にマークアップすることが難しいという点も事実である．

しかし，国際標準に則った基盤テクストを構築し，国際的に流通させることは，世界における日本の研究および東アジアの研究にとっても重要である．単に全文が Web 上にあるのではなく，構造化されたテクストとすることで，『延喜式』そのものを直接読まなくとも，なんらかの見当をつけるなどの使用も考えうる．日本を対象とする歴史研究者のみならず，中国を対象とする研究者などに対しても有益になりうる．というのも，『延喜式』は，古代の法律である「令」の施行細則であるためである．中国の律令を日本が受容しそれをどのように展開させたかを知るための重要な資料であるとともに，いまは失われている唐代中国の法令関係のヒントとなりうるものを含むなど，東アジアにおける研究価値は高い．これらの点からも，より汎用的なマークアップを施すことが必要であると考える．

そこで，テクストをどのように理解したかを書く TEI に対し，さらにメタなレベルでのマニュアルを作成することで，TEI データを多くの人が基盤データとして作成可能にすることを目指した．

さらに付け加えるなら，万一，この延喜式 TEI データがなんらかの理由で歴博から離れて管理されることになっても，このマニュアル自体が，当時どのような意図で構築されたかの記録となり，長期的なデータ活用につながりうると考える．データの長期的保存を行う際には，データそのものの保存形式，媒体，フォーマットなども課題になるが，いざ残ったものを閲覧した際に，「これはどのような意図で作られたものなのか」がわからず，結果的に使えないという事態が起こる．データを作った際の，意図・精度・大元として参照した資料などの情報がないものは，結果的に使えずに廃棄される可能性が高くなる．

そのような観点からも，データと合わせてこのような記録を作ることが重要であると考えた．たとえ，オープンデータであっても，単にオープン

で漂流したデータは，再度それ自体の価値を問い直す必要がある．

4．マニュアルの具体的構成および内容

次に，本マニュアルの具体的な説明に移る．本TEIマニュアルの具体的な構成は下記の通りである．

1. メタデータ記述

ここには，基本的には TEI Header を中心に説明を述べている．資料について説明する必要な要素・データ作成・資料作成などに関連した人物，画像との対応付けなど，歴史的な資料をエンコーディングする際に，汎用的に必要であろうと考えられる部分に関する基礎的な説明を述べている．

titleStmt
　title：史料名
　editor：辞書上の編纂者名
　respStmt：著者や編集者など特定の役割を示す要素が充分ではない場合に，テキスト，版，記録などの知的内容に関する責任を示す
　　resp：人物の知的責任の性質を表す一節を示す
　　persName：@ref="(VIAFのID)" を用いて国際人名典拠ファイルとリンク付け

例

```
<respStmt>
  <resp>TEIマークアップモデルの検討と実装</resp>
  <persName>小風尚樹</persName>
</respStmt>
```

editorialDecl：電子テクストのマークアップにあたっての方針などを記述

図２　メタデータ記述のマニュアル

また，ここには TEI Header の後に記述する画像との対応づけに関する部分もマニュアルとして記載することとしている．IIIF と TEI との連携は，コンテンツの検索・閲覧・国際標準という点では大変に相性が良く，著者らのうち小風・永崎がすでにプロトタイプを実験的に発表している[9].『延喜式』についても，画像での公開も検討されており，それらのデータとともに本マニュアルが公開されることで，同様の試みが広がるのではないかと考えられる．ただし，このような作業というよりは研究に属するものを，汎用的な説明として入れるべきか，それとも，研究的な事例として入れるべきかなどは，このマニュアルの具体的な読者を想定しつつ，作成する必要がある．

＊画像との対応付け
・史料画像とテクストの対応関係を記述したい場合の方法として、ここではOxygen XML Editorに搭載されているImage Map Editorの機能の使い方を紹介する。

・使用するタグ

facsimile	teiHeaderの終了タグの直後に配置する
surface	画像一枚ごとの情報をまとめるためのタグ
graphic	URLや相対パスで画像を指定するタグ（@url属性を用いる）
zone	Image Map Editorにより@xml:id付きで自動生成されるタグ

・Image Map Editorについて
Oxygen XML Editorには3つのモードがあり、それぞれ「テキスト」「グリッド」「作者」である。そのうち「作者」モードは、マークアップしたテクストを簡単に変換して、見やすい形で表示してくれる機能を持っているのだが、画像との対応付けを符号化するImage Map Editorもこの「作者」モードに含まれている。

上記のようにgraphicタグで画像のURIを指定すると、「作者」モード中のImage Map Editorで、画像のどの部分を切り取って座標情報をデータ化するか指定することができる。画像の切り取りにあたっては、長方形での切り取り・多角形ポリゴンの切り取りを選択することができる。下図は、国立国会図書館デジタル・コレクションでIIIF形式で提供されている延喜式画像の一行ごとの座標情報を切り出したものである。キャプチャ右側のID欄に見られるように、@xml:idが機械的に生成されている。

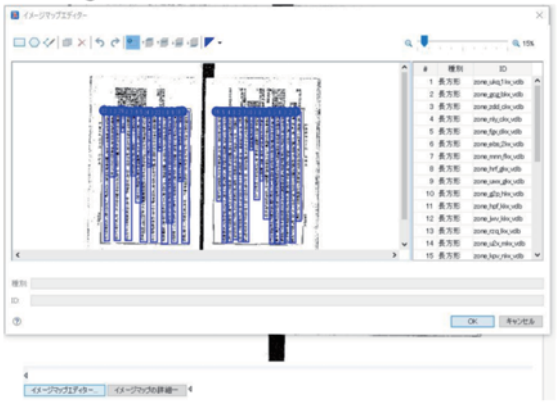


図３　Image Map Editor などの解説例

2. 全体構造記述：巻や章など，区切りごとに構造化することの必要性

『延喜式』の特性に応じて，どの部分にどのようなタグを付したのかを説明している．この部分

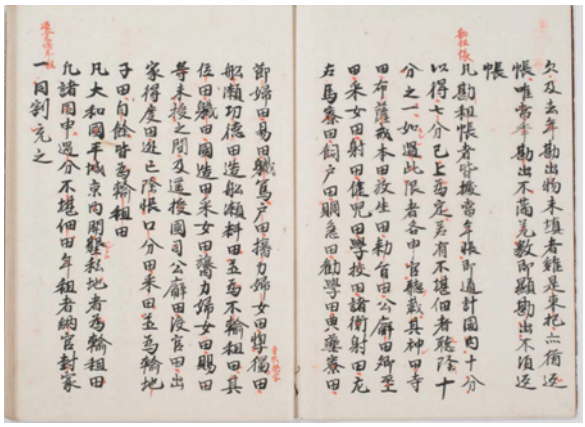


図４　『延喜式』のうち「文章型」のものの例

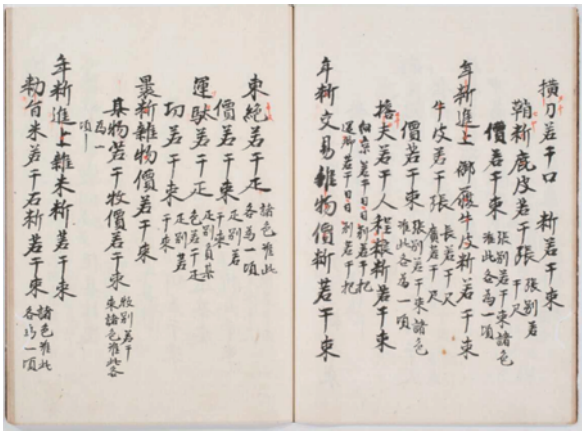
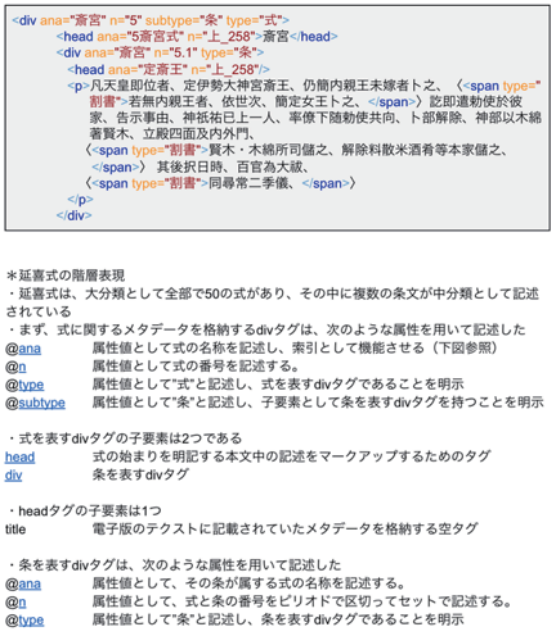


図５　『延喜式』のうち「帳簿型」のものの例

は一般的には資料の特性に応じて変更しなければならない部分ではある．ただし，『延喜式』には大きく２つ，もしくは３つの文の様式を持っている．一つは，祝詞などのような漢文の文章体の様式，もう一つは法令ごとに帳簿のように説明を記した帳簿様式である．さらに分けるなら，文書の例示のような「見本」様式が帳簿様式から分離できる．このように，『延喜式』は，複数の様式を持っており，比較的多くの資料でも参照しやすい特徴がある．そのため，『延喜式』を例とすることで比較的汎用性の高いマニュアルとなると考えられる．



＊延喜式の階層表現
・延喜式は、大分類として全部で50の式があり、その中に複数の条文が中分類として記述されている
・まず、式に関するメタデータを格納するdivタグは、次のような属性を用いて記述した
@ana　属性値として式の名称を記述し、索引として機能させる（下図参照）
@n　属性値として式の番号を記述する。
@type　属性値として"式"と記述し、式を表すdivタグであることを明示
@subtype　属性値として"条"と記述し、子要素として条を表すdivタグを持つことを明示

・式を表すdivタグの子要素は2つである
head　式の始まりを明記する本文中の記述をマークアップするためのタグ
div　条を表すdivタグ

・headタグの子要素は1つ
title　電子版のテクストに記載されていたメタデータを格納する空タグ

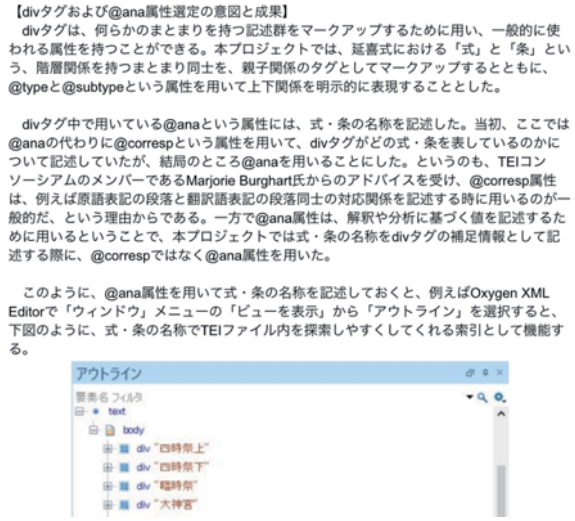
・条を表すdivタグは、次のような属性を用いて記述した
@ana　属性値として、その条が属する式の名称を記述する。
@n　属性値として、式と条の番号をピリオドで区切ってセットで記述する。
@type　属性値として"条"と記述し、条を表すdivタグであることを明示

図６　『延喜式』の階層構造部分のマニュアル

3. 目的に即した記述

これは著者のうち小風が検討したものなど，関連するマークアップを記録として残したものが，現時点では入っている．ここには，『延喜式』マークアップのための個別の研究で行われたデータを蓄積する．これ自体は，必ずしも汎用的ではないが，基盤研究から発展した検討を行う際，どのようなことができるのかの参照を行うことを目指している．

特に，小風の研究においては，鮑を中心とした，トランザクショノグラフィの研究である．また，それに関連して，『延喜式』上にある諸国の財政を米をベースとして検討する実験などを実施している．これらの研究自体は，実験段階でもあるものの，いくつかの研究上の示唆を与えるものであったとともに，『延喜式』における資料的な限界などをこれまで研究で指摘されていた部分を改めて，誰でもがトレースできる形式でコンピュータで証明したことにも重要な意義がある．これらの研究に関係する情報について，具体的にどのようにマークアップしたかなどがこの章には記載されている．



【divタグおよび@ana属性選定の意図と成果】
divタグは、何らかのまとまりを持つ記述群をマークアップするために用い、一般的に使われる属性を持つことができる。本プロジェクトでは、延喜式における「式」と「条」という、階層関係を持つまとまり同士を、親子関係のタグとしてマークアップするとともに、@typeと@subtypeという属性を用いて上下関係を明示的に表現することとした。

divタグ中で用いている@anaという属性には、式・条の名称を記述した。当初、ここでは@anaの代わりに@correspという属性を用いて、divタグがどの式・条を表しているのかについて記述していたが、結局のところ@anaを用いることにした。というのも、TEIコンソーシアムのメンバーであるMarjorie Burghart氏からのアドバイスを受け、@correspの属性は、例えば原語表記の段落と翻訳語表記の段落同士の対応関係を記述する時に用いるのが一般的だ、という理由からである。一方で@anaの属性は、解釈や分析に基づく値を記述するために用いるということで、本プロジェクトでは式・条の名称をdivタグの補足情報として記述する際に、@correspではなく@anaの属性を用いた。

このように、@anaの属性を用いて式・条の名称を記述しておくと、例えばOxygen XML Editorで「ウィンドウ」メニューの「ビューを表示」から「アウトライン」を選択すると、下図のように、式・条の名称でTEIファイル内を探索しやすくしてくれる索引として機能する。

図７　階層構造のうち属性の理由などを記述した部分

span　本プロジェクトでは、@type="割書"として、割書きの註をマークアップした

persName　teiHeader内などで@xml:idを付与したpersNameタグの属性値を@ref属性を用いて参照する
placeName　延喜式に出てくる地名を、@xml:idと紐づけてマークアップ
roleName　延喜式に出てくる職名を、@xml:idと紐づけてマークアップ
orgName　延喜式に出てくる組織名を、@xml:idと紐づけてマークアップ
measure　種々の品目をマークアップする（詳しくは後述）
num　数量をマークアップ
【以下のタグは、近々ガイドラインのアップデートで入る予定】
unit　様々な単位をマークアップ
unitDecl　encodingDesc内でunitDefタグを格納するカスタマイズタグ
unitDef　unitDecl内で単位の定義を格納するカスタマイズタグ
【以下のタグについては、Tomasek & Bauman（2013）、小風訳（2015）を参照】
hfr:listTransaction　hfr:transactionを格納するためのカスタマイズタグ
hfr:transaction　hfr:transferを格納してモノの移動を記述するカスタマイズタグ
hfr:transfer　モノの移動そのものの情報を記述するカスタマイズタグ

図８ 目的に即した記述のマニュアル　特にカスタマイズタグも用いているので，その理由なども記している．

今後も，小風以外にも『延喜式』のマークアップを用いた個別研究事例をここに蓄積する予定である．

また，これ自体は研究の記録としても機能し，一つの資料に対してどのような研究が行われた

のか，実際に流通しているマークアップデータはどのような意図で作られたのかを残すものでもある．このようなデータは，一義的には論文で記述されるが，それをより具体化したものをここに残すことで，より高度な研究への貢献を期待できる．

4. 多言語対応

『延喜式』のプロジェクトにおいては，条文を漢文のみならず，読み下し文，現代語訳とともに，英訳することまでスケジュールの中に入っている．しかし，英訳の作業自体は，どのような形式で英訳を可能にするか，全体を英訳するのかダイジェストかするのかなど，まだ議論すべき点が多くある．そのため，現時点では中途であるため，まだマニュアル上では記述されていない．今後のプロジェクトの進捗に応じ，充実すべき箇所である．さらにいえば，漢文（白文）・読み下し文・現代語訳・英訳を今後，どのように検索し，表示させるかというユーザインターフェースの重要な課題ともセットになってくる．複雑な構成を持つ『延喜式』の中で，どのようなデータの提示が望ましいのか，そしてどのようなデータ形式を持つのが望ましいのかを検討した上で，その検討の考え方とともに・書かれることになるであろう．

5. スキーマ

TEIのデータをどのようにカスタマイズしたかの記録である．長期的なデータ活用のためでもある．

6. 表示やアプリケーションの例

ここで作成したデータが，どのように応用されるかの例を記載している．最終的な表示方法や，アウトプットも含め，ここに記載している．

全体の構成としては上記の通りである．主に 1・2 で最も基本的なマークアップを可能にし，3・4・5 においてより応用度の高いものを示すという構成となっている．6 はそれらの流れと少し異なり，TEI の意義を示すための機能も果たしている．たとえば，1 で示したような，画像との対応づけや，多言語対応の結果としての表示方法なども，ここで表現されることになる．

５．本マニュアルの意義と位置づけ

本マニュアルを使う対象については，以下のように想定する．

a. 『延喜式』の基盤データ作成者

b. 日本を中心とする東アジア前近代資料の TEI データ作成を行う研究機関・資料所蔵機関担当者

c. TEI によるテクストデータを活用し，研究を行う人文情報学研究者

d. 現在ではないが，本プロジェクト終了後，データが長期的に流通した際，後にデータ分析を行う人物

a および b については，マニュアルの 1 と 2 を読んでもらうことを意図している．

『延喜式』は全部で 50 巻のテクストである．そのため，ある程度までは機械的に対応が可能であったとしても，その先は人間によるチェックが欠かせないものとなる．その際，このマニュアルを活用し，データ化を行うことを想定している．また，東アジア資料の TEI データ作成という観点からは，特に日本におけるテクストデータの新たな起爆剤となりうることができればと考えている．

TEI は，非常に有益な手法であること自体は，広く認識されているものの，実際にはエレメントの多様さやルールの複雑さによって，忌避されることが多いものとなってしまっている．一方で，研究者が進める場合には，比較的明確なアウトプットとともにでなければ，作業をするのが難しいという実情がある．そのため，研究者そのものではなく，研究補助者が基礎的な作業を行い，研究者はその中でも研究のためのマークアップなどの検討する必要があると考えられる．そのような場合，このような基礎マニュアルは有効であろう．

c については，おそらくは 3 以降を必要に応じて読むことになるであろう．従って，厳密にはマニュアルそのものを参考にするというよりは，研究論文をトレースしたり，事例の参考にするといった利用方法が考えられる．

まずは，2 までを共有することで，より多くのデータが構築されることが，重要である．基礎データを蓄積し，その上に研究上のマークアップがなされることで，TEI の意義はより深まってくるものであろう．さらにいえば，このような作業のメタな記録を整備し流通させることで，より広範なデータの蓄積へとつながりうるのではないかと考える．

６．意義と課題

本マニュアルの作成は，緒についたばかりである．そのため，今すぐに直接的な意義を語るのは

困難な部分がある．しかし，このような「メタのメタ」的な作業の蓄積は，人文情報学や「デジタルアーカイブ」にとっては，行う必要のある作業である．多くのデータベースが生み出されている中で，長期的なデータ維持を行えているのは，「単にその機関が維持していて，維持する予算を確保できているから」に過ぎないところが多い．無論，その中で，力のある人文情報学者がデータの「棚卸し」を行い，新たな形に衣替えするところもある．しかし，流動性が高くなっていく今後において，このような状況が維持できるという保証はない．事実多くの時限的プロジェクトに支えられた「デジタルアーカイブ」は消えていく運命にあるものが多かったのも事実である．

その意味では，いかにデータとその関連記述を効果的に残し続けていくかは，今後の未来にとっては重要であろう．

課題としては，このマニュアル自体が，「では，読んでわかるようなものになるのか」という観点にある．現在は，ある程度プロジェクトを理解した人物が作成している．そのため，今後は，これらのマニュアルを実地で活用したり，他者のレビューを受けたりしつつ，「わかる」マニュアルの構築を行わなければならないであろう．

7.おわりにかえて

日本におけるテクストデータなどの流通と長期的展開は，極めて重要な課題である．今後の人文情報学に関わるデータの充実のためにも，このようなマニュアル整備が一般的になり，かつ流通することを強く望む次第である．本研究が，その端緒となれれば幸いである．

参考文献

1）虎尾俊哉『延喜式』（吉川弘文館，1964 年）

2）桶谷猪久夫, Delmer Brown, 大久保祐子, 山尾正之 : XML を利用した日本古典史料の英日全文連携検索システムの構築. 国際研究論集. 19 (1), 87-110. 2005.

3）Kathryn Tomasek and Syd Bauman: Encoding Financial Records for Historical Research. Journal of the Text Encoding Initiative [Online] 6. 2013. 入手先〈https://journals.openedition.org/jtei/895〉（参照 2018-09-07 以下同）．小風による和訳 〈http://hdl.handle.net/2261/56940〉

4）Naoki Kokaze, Kiyonori Nagasaki, Makoto Goto, Yuta Hashimoto, Masahiro Shimoda, Albert Charles Muller, : TEI/XML Markup of Engi-shiki as Research Platform for Historians of Ancient Japan, The Japanese Association for Digital Humanities 2017, Kyoto. Sep 2017.

5）Naoki Kokaze, Kiyonori Nagasaki, Makoto Goto, Yuta Hashimoto, Masahiro Shimoda, and Albert Charles Muller: TEI/XML Methodological Examination on Unit Conversion not based on the Metric System. The 2017 Annual Meeting of the TEI Consortium, Victoria, Canada. Nov 2017; How to encode measurement, 入手先〈https://github.com/TEIC/TEI/issues/1707〉.

6）高橋洋成, 永井正勝, 和氣愛仁, : 画像，TEI，LOD を用いた文字研究・言語研究のためのプラットフォームの構築 情報処理学会研究報告 2015-CH-105(5),pp1-8. など

7）永崎研宣 : 仏教文献のための構造的なデジタルテクストの記述と活用. 印度學佛教學研究 63 (2): 1094-1088. 2015.

8） 図書館向けのマークアップマニュアルとしては以下のようなものもあるが，さらに研究機関向けのものを目指す．〈http://www.tei-c.org/SIG/Libraries/teiinlibraries/3.1.0a/main-driver.html〉

9）Nicholas Laiacona, Ben Brumfield, Naoki Kokaze, Kiyonori Nagasaki, Makoto Goto: Connecting TEI and IIIF. 2018 IIIF Conference, Washington DC. May 2018. 入手先〈https://iiif.io/event/2018/washington/program/paper-61/〉

謝辞

本研究プロジェクトは、人間文化研究機構　基幹研究プロジェクト（広領域）「異分野融合による「総合書物学」の構築」および人間文化研究機構機関拠点型基幹研究プロジェクト「総合資料学の創成と日本歴史資料のバックアップ」の成果の一部である。また、「総合書物学」の研究に関わり、国立歴史民俗博物館の小倉慈司氏、清武雄二氏には多大なるご教示をいただいた。記して御礼を申し上げる。