大正新脩大蔵経の構造的記述に向けて

渡邉要一郎（東京大学）
永崎研宣（人文情報学研究所）
朴賢珍（東京大学大学院）
王一凡（東京大学大学院／人文情報学研究所）
村瀬友洋（大蔵経研究推進会議）
渡邉眞儀（浄土宗総合研究所）
大向一輝（東京大学）
下田正弘（東京大学）

SAT 大蔵経テキストデータベース研究会は文字にして漢字 1 億字超に及ぶ膨大なテキストデータを所持し，公開しているものの，いまだその電子テキストは十分には構造化されてきていなかった．そこで，本研究会は Text Encoding Initiative に従って，この仏典電子テキストの構造化を行っている．本論文では，予想される大規模な人員による作業のために当研究会で検討された作業手順と，大正大蔵経マークアップの方針が論じられる．

Structural Description for the Taisho Tripitaka

Yoichiro Watanabe (The University of Tokyo)
Kiyonori Nagasaki (International Institute for Digital Humanities)
Hyunjin Park (Graduate school of the University of Tokyo)
Yifán Wáng (Graduate school of the University of Tokyo, International Institute for Digital Humanities)
Tomohiro Murase (Council for the Promotion of Tripitaka Research)
Masayoshi Watanabe (Jodo Shu Research Institute)
Ikki Ohmukai (The University of Tokyo)
Masahiro Shimoda (The University of Tokyo)

Although the SAT Daizōkyō Text Database Committee possesses and makes available to the public a vast amount of textual data containing over 100 million Chinese characters, these electronic texts have not yet been sufficiently structured. The present study aims to structure the e-texts according to the Text Encoding Initiative. This paper discusses the reviewed work procedures for the expected large-scale manpower to be employed and the policy for marking up the e-texts.

１．まえがき

SAT 大藏經テキストデータベース研究会（以下，当研究会とする）は大正新脩大蔵経（以下，大正蔵とする）の 2920 件の仏典，文字にして漢字 1 億字超に及ぶ電子テキストデータを保持し，このデータを公開している[1]. その利便性をさらに高めるため，本研究会は現行の電子テキストを TEI ガイドラインに準拠して構造化するべく準備を進めている．これだけの膨大なテキストの統一的に構造化は，かなりの大規模な作業となる．したがって，その大規模作業に着手する前に，マークアップの方針と，具体的な作業手順の方針を策定しておくことは必須である．というのも，TEI ガイドラインは当初からグローバルなものを志向していたが，当初の参画者のほとんどが西洋世界の研究者であったために，英語を中心とした西洋文献が主要な対象になってしまっている現状がある．そのため，基本的には既存の TEI ガイドラインに準拠しつつも，不足分を補うようなかたちで漢文や日本語文献のためのマークアップ方針をまず策定する必要がある．

当研究会のマークアップ方針は，大正蔵の電子上での正確な再現である．大正蔵の本文には誤りも少なくないし，表記の不統一も存在する．それをあえて統一することなく，仏教研究の基盤とし

て使用され続けてきた大正蔵をできる限りそのままの形で提供することを現時点での目標としている．これは大正蔵そのものが研究の道具というだけでなく，研究対象にならざるを得ないという側面も有しているからであり，また現状で研究の基盤となっている大正蔵と異なるヴァージョンのテキストがウェブ上に存在することによって，研究上いたずらな混乱や，過去の研究とのあいだに無用な断絶が生ずることを避けるためでもある．

本稿は，そのような大規模作業の準備段階の一例，また，漢文資料のマークアップ方針に関する一例を提示する．

２．従来の状況

当研究会は，大正蔵のタグ付き電子テキストを保持しているが，これは 1994 年に作成されたものであり，これは XML (1998)よりも古い．したがって，TEI/XML 化のためには抜本的な書き換えと追加すべき数多くの情報が存在する．また，SAT と同じく大正蔵の電子テキストをウェブ上で公開している中華電子佛典協會（以下 CBETA）は，TEI を独自拡張した XML ファイルを公開している[2]．しかしながら，CBETA に見られるような大幅な拡張は必ずしも必要ではなく，従来のタグと属性の組み合わせで相当程度目的を果たすことが可能である．本研究会は，独自拡張したタグを可能な限り避け，現状の TEI ガイドラインのなかで完結する方針をとっている．これにより，例えば Oxygen 等の既存のエディターを用いて作業する際にもエラーが起こりにくくなり，作業者の負担を軽減することにつながる．さらにデータの可読性と汎用性も，より改善されるようになると思われる．CBETA のマークアップには，更なる細かな課題が存在するが，これに関しては 5 節で言及する．

３．作業の形態

当研究会では，コロナ禍の以前はおおむね週に一回程度，主に仏教研究を専門とする作業分担者が実際に人文情報学研究所に集まり，具体的なマークアップを暫定的に試み，それを修正しつつ作業を行っていた．現在の作業担当者は 6 名程度であるが，1 節に先述したように，将来的には大人数での作業が必要となる．その際の作業担当者は，内容の理解が必要になるため，漢籍の扱いにある程度の習熟を前提とするが，情報技術に対する高度な前提知識は要求できない．したがって，大規模作業の際にはある程度マニュアル化され，機械的に適応できるようなマークアップ方針が準備されている必要があり，現在はその策定を行っている．

個々人の作業結果は複数人による確認や修正が行われる必要があるため，共有されなければならない．そのため，当初は Google Drive によるファイル共有を行っていたものの，主に差分管理の面で不都合が生じる場合が多く，GitHub を利用することになった．この際，GitHub 使用に関するマニュアルを作成し，講習を行うのみならず、講習を通じたマニュアルの検証も実施し，将来の作業担当者が GitHub の使用に適切に習熟できるような仕組みを整えている．

４．大正蔵の三分類

大正蔵は目録部・図像編を除くと、印度撰述部・中国撰述部・日本撰述部に大別される．印度撰述部・中国撰述部に含まれる仏典の大部分は，13 世紀に刊行された高麗版大蔵経再雕本や，万暦版大蔵経等といった過去の大蔵経に入蔵されており，その時点である程度フォーマットが整備された．したがって，それを継承した大正蔵でもそれほど大きな版型の揺れは無いものと想定された．一方，日本撰述部に関しては，初の入蔵となるテキストがほとんどであり，また，写本や流布版本など，様々なフォーマットの資料から直接書き起こされたためにフォーマットが多様であり編集方針も版型も前者と比して統一性が見られない．そのため，当研究会では差し当たって印度撰述部・中国撰述部に関する方針を策定し，その後それを応用する形で日本撰述部の作業を行うこととした．

５．印度・中国撰述部構造化の問題

本節が主題とするテキストは印度撰述部と中国撰述部である．

大正蔵に含まれるテキストは章立て等の内容の区分とは必ずしも一致しない「巻」という単位で区切られる場合がほとんどである．この「巻」は fascicle「巻物」の謂いである．刊本となった時代の大蔵経の上にも，かつてテキストの形態が巻物であった時代の痕跡が残り続けているわけである．しかし，この「巻」の区分は例えば注釈家がテキストを参照する際に用いていることもあり，研究にとって有意味であって蔑ろにされるべきではない．

各巻の冒頭にはテキスト名・著者・訳者等の，テキスト外部情報が記述されており，多くの「巻」の末尾には，さらにタイトルが付与されている場合が多い．このような「巻」による区分の仕方は，大正蔵の版型にも反映されており確認は容易で

ある．しかしながら，「巻」は物理的な区別であって，内容上の区別を必ずしも意味しておらず，長さはおおむね一定である．

一方で，意味上の区別として「品」が存在している．これは chapter「章」を意味するもので，物理的な長さは特に限定されておらず，著者の判断により長短が生じている．この「巻」と「品」は異なる次元に存在する概念である．つまり，意味上の区分である「品」が，巻物というテキストが記述された物理的形態の限界によって，複数の「巻」に跨って存在することは容易に起こりえるのである．

このような内容上の次元と，物理的な次元の区別は，従来のタグ付きテキストではあまり意識されてこなかった点であり，方針策定にあたり特に整備が必要であった．

６．マークアップ方針

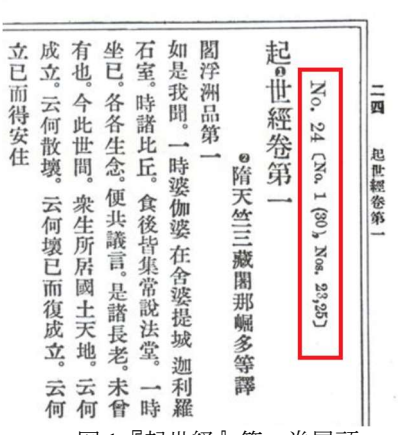
このような状況を踏まえ，現段階で当研究会が規定した構造について以下に記す．先にマークアップファイルの骨格を概略すると以下の通りである．

```
<TEI>
    <teiHeader> … </teiHeader>
    <text>
        <body>
            <div type="taisho_head">
                No. xxxx（大正蔵番号）
            </div>
            <div type="taisho_body">
                本文
            </div>
        </body>
        <back>
            <note>...</note>
            <note>...</note>
            <applist>
                <app>...</app>
                <app>...</app>
                .....
            </applist>
        </back>
    </text>
</TEI>
```

以下，順次項目について説明する．

■ について

この箇所に記述するべき箇所は大正蔵本文開始前の図 1 の四角で囲まれた，明らかに大正蔵編者によって記述された部分である．


二四　起世經卷第一

No. 24 (No. 1 (30), Nos. 23,25)

起世經卷第一

隋天竺三藏闍那崛多等譯

閻浮洲品第一

如是我聞。一時婆伽婆在舍婆提城迦利羅石室。時諸比丘。食後皆集常說法堂。一時坐已。各各生念。便共議言。是諸長老。未曾有也。今此世間。衆生所居國土天地。云何成立。云何散壞。云何壞已而復成立。云何立已而得安住

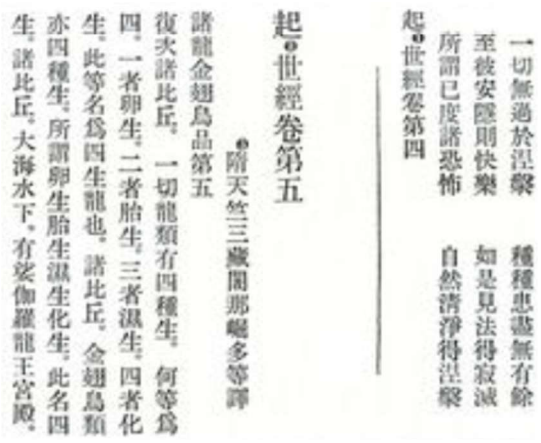
図 1『起世経』第一巻冒頭

大正蔵編者によって書かれた部分は経文と区別されるべき事項であるから、なんらかの方法で区別を示す必要がある．漢語仏典のテキスト番号を示すインデックスとして No. 24 等の番号は研究において広く用いられているため，この情報は必要であり，大正蔵内の位置情報と合わせて記述される必要がある．図 1 の例は次のように記述される．

```
<div type="taisho_head">
    <p><lb n="T0024_.01.0310a01"/>
    No.24[No.1(30),Nos.23,25]
    </p>
</div>
```

■ について（1）

 タグ内は，いわゆるテキスト本文が記述される．本文記述の一例として，図 2 に『起世経』という経典の五巻冒頭を挙げた．またその直下にこれのマークアップ例を記している．


一切無過於涅槃　種種患盡無有餘　至後安隱則快樂　如是見法得寂滅　所謂已度諸恐怖　自然清淨得涅槃

起世經卷第四

起世經卷第五

隋天竺三藏闍那崛多等譯

諸龍金翅鳥品第五

復次諸比丘。一切龍類有四種生。何等爲四。一者卵生。二者胎生。三者濕生。四者化生。此等名爲四生龍也。諸比丘。金翅鳥類亦四種生。所謂卵生胎生濕生化生。此名四生。諸比丘。大海水下。有婆伽羅龍王宮殿。

図 2『起世経』巻第五冒頭

```
<milestone unit="fascicle_beginning" n="5"/>
<div type="chapter" n="5">
    <ab type="fascicle_beginning">
        <title type="fascicle_beginning">
        起世經卷第五
        </title>
        <persNamerole="translator"
        ref="http://viaf.org/viaf/110570907">
        隋天竺三藏闍那崛多
        </persName>等譯
    </ab>
    <p><title>諸龍金翅鳥品第五</title></p>
    <p>諸龍金翅鳥品第五の本文</p>
</div>
```

5 節に前述したように，本文の意味上の区分である「品」を区切る場合には ，「巻」の変更地点を表示する <milestone>,「巻」に従属する <ab type="fascicle_beginning"> 部分の階層は上記のように表現できる．先述のように，従来の当研究会が保持していたタグ付きテキストは，「巻」の単位を タグで区別しているだけであり，内容を意味する「品」の単位には，ほとんど手がつけられていなかった．

このような細分化されたマークアップによって，例えば純粋な本文のみの抽出，品単位による抽出，巻単位による抽出など，研究者の求める情報がより容易に得られるようになるものと思われる．

そのなかでさらに，「品」に対応する範囲を，テキストの区分を意味する 等のタグで区別することにし，巻の切り替えには，上記のようなテキスト構造上の変化を意味しない <milestone unit="fascicle_beginning" n="1"/> というタグと属性を用い，頁や行の変更の変更と同等のものとして扱おうと考えた．

XML のタグは例えばパラグラフを明示する タグに対して， のような，該当パラグラフの終了を示す閉じタグが対応していなければならないが，頁の区切りや行の区切りといった要素は，タグのあいだに囲むものが存在しえない．そのような場合，<pb/>, <lb/> といった，それぞれ頁・行の区切りを明示するだけの，挟まれる要素を持たない「空タグ」が用いられる．「巻」の区切りを表示するために用いた <milestone/>タグの用法もこれらに準じている．

また，各巻冒頭には，訳者・タイトル等の情報が付与されている．これに対してはパラグラフ内テキストの特殊な部分を表示する <ab type="fascicle_beginning"> によってマークアップを行った．このブロックのなかには経典名のほかに，筆者・訳者の情報が含まれる．これに対しては <persName> タグによるマークアップを行っており，ref の属性によってパーマリンク（http://viaf.org/）を示すことが可能になっている．加えて，role の属性によって，その人物の該当テキストにおける役割を示している．

また，このブロックは「巻」単位に付属する部分なので，「品」の意味単位のみに注目する場合は，この部分のみを選択的に無視してテキストを抽出するなどの処理を可能にするであろう．

■ について（2）

前述したように，意味上の区分たる「品」と物理的区分たる「巻」は必ずしも一致しない．『起世経』の一例はその好例となろう．

図 3 は『起世経』二巻途中から始まる「地獄品」と巻三，四の冒頭を大正蔵本文から抜き出したものである．ここで「地獄品第四之一」「地獄品第四之二」「地獄品第四之三」とあるものは，subchapter とはいえないものであり，意味の区切りたる「地獄品」が，「巻」の物理的制約のもと分かたれてしまった結果生じた，二次的な区分けであると理解できる．本研究会はこの構造を以下のようにマー

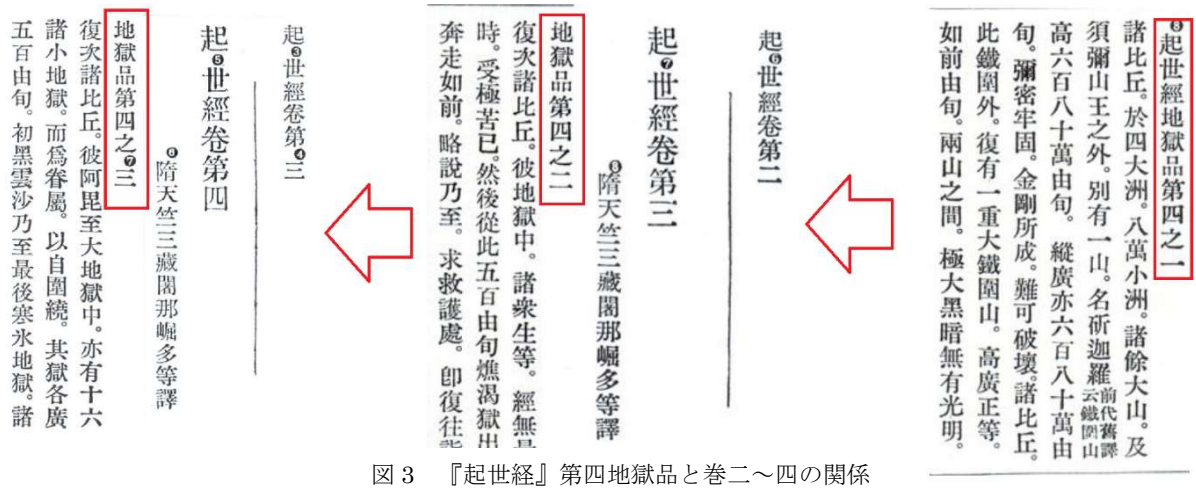


図 3　『起世経』第四地獄品と巻二～四の関係

クアップする（<ab>タグ内は省略）.

```
<milestone unit="fascicle_beginning" n="2"/>
……
<div type="chapter" n="4">
    <p>
        <title type="subdesc"
        subtype="physical"n="4.1">
            起世經地獄品第四之一
        </title>
    </p>
    <p>
        『起世經』地獄品第四之一の本文
    </p>
    <title type="fascicle_end">
        起世經卷第二
    </title>
<milestone unit="fascicle_beginning" n="3"/>
    <ab type="fascicle_beggining">
        <title>
            起世經卷第三
        </title>（省略）
    </ab>
    <p>
        <title type="subdesc"
        subtype="physical" n="4.2">
            地獄品第四之二
        </title>
    </p>
    <p>
            『地獄品』第四之二の本文
        </p>
        <title type="fascicle_end">
            起世經卷第三
        </title>
<milestone unit="fascicle_beginning" n="4"/>
    <ab type="fascicle_beggining">
      <title>
          起世經卷第四
      </title>（省略）
    </ab>
    <p>
        <title type="subdesc"
        subtype="physical" n="4.3">
            地獄品第四之三
        </title>
    </p>
    <p>
        『地獄品』第四之三の本文
    </p>
</div>
```

このように，二次的なタイトルは挿入されているものの， タグで区切らず，subtype="physical" 属性によってこれが「巻」による物理的限界によって区分された，いわば見かけ上の subchapter であることを示している.

CBETA は「地獄品」全体を<cd:mulu level="1">とし，これら「見かけ上の subchapter」に対して<cd:mulu level="2">を振りあて，あたかも真なる subchapter のように表記しているが，意味上から真に subchapter とすべき場合と，およそ意味区分が想定されていなかったであろう単位が，同一のレベルと混合される懸念があり，改善の余地があると思われる.

■ <back> について

<back>タグ内部には，異読情報や校訂に際して使用したテキストに関する情報等のうち，大正蔵の版面には記述されているものの，大正蔵の本文とは見做されない要素を記述している.

大正蔵の脚注には異読情報とそうでない編者のノートが混在している．これも可能な限り大正蔵の内容を再現しつつ，注釈の種類による分割を試みた.

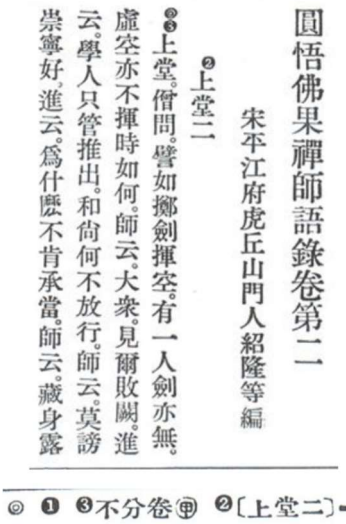

圓悟佛果禪師語錄卷第二
宋平江府虎丘山門人紹隆等編
❷上堂二
❶❸上堂。僧問。譬如擲劍揮空。有一人劍亦無。虛空亦不揮時如何。師云。大衆。見爾敗闕。進云。學人只管推出。和尚何不放行。師云。莫謗崇寧好。進云。爲什麼不肯承當。師云。藏身露
◎ ❶ ❸不分卷㊥ ❷〔上堂二〕一㊥

図 4『圓悟佛果禪師語録』第二巻冒頭とその脚注

一例として，上記テキストにおいて，❷は「甲本」においては消去されているという意味の異読情報であることが，大正蔵の省略記法を用いて表記されている一方で，❸は「『甲本』においては巻が分けられていない」ということをそのまま漢文で書いている編者の注記となる．❷は一般の critical apparatus を表記する 3 つの記法のうち、double end-point attachment method [3]に従って

```
<app from="#tft_0718_2" to="#tft_0718_2e">
    <lem wit="#大正">上堂二</lem>
    <rdg wit="#甲"/>
</app>
```

と表記される．このとき，本文中の適当な位置がアンカータグ <anchor xml:id="tft_0718_2"/>, <anchor xml:id="tft_0718_2e"/> によって囲まれており，注釈対象文字列が明示されている．❸のような表現はこのままでは表記しにくい．当研究会では，critical apparatus のリストの外部に note タグを独立させ，そこに❸のような情報を記すことにした．❸の例は

```
<note target="#tft_0723_1" type="footnote">
  ◎不分卷＜甲＞
</note>
```

と表記される．これにより，大正蔵の脚注表記をそのままの形で残すことが可能になる．

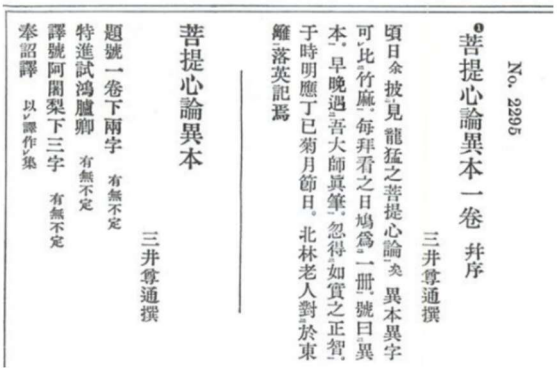
７．日本撰述部について

日本撰述部のテキストは，CBETA においては公開されておらず，当研究会独自に公開されているデータとなっており，データ提供の貴重性という観点からみても重要である．

日本撰述部に関しては，本稿 5 節に先述した通り事情が異なり，複数のフォーマットが存在している．

まずは実際にどのような構造があり得るのかを統一的に把握する必要があった．そのため，作業分担者が個々のテキストを実見し，構造のあり方を調査し，個々人が Google スプレッドシートにそのテキストがいかなる構造であるかを記入していった．

その結果，日本撰述部のテキストには，構造があらかじめ序文等で明示され，明らかに構造化が意図されたものが多いことが分かった（一例として図 5 参照）．


No. 2295

❶菩提心論異本一卷 幷序

三井尊通撰

頃日余披見龍猛之菩提心論異本異字可比竹麻。毎拜君之日焉爲一冊。號曰異本。早晩遇吾大師眞筆。忽得如實之正智。于時明應丁巳菊月節日。北林老人對於東離落英記焉

菩提心論異本

三井尊通撰

題號一卷下兩字 有無不定

特進試鴻臚卿 有無不定

譯號阿闍梨下三字 有無不定

奉詔譯 以上譯作ノ集

図 5 『菩提心論』冒頭

さらに，前述の「巻」の事情とは異なり，日本撰述部の「巻」は，内容上の区分とより密接しているものも多く，分量もテキストごとに隔たりが大きい．各巻の冒頭にその巻で扱う内容の梗概を提示するテキストも数多く見受けられた．印度・中国撰述部で策定した基本的な方針に依拠しつつ，日本撰述部における特殊な部分に対する個別のマークアップ方針が必要であろうと思われる．

また，目録形式・儀礼の式次第等の印度撰述部・中国撰述部では見られない種類の構造をもったテキストが若干の割合で含まれ，これらに対するマークアップ方針は現在検討中である．

８．まとめ

以上のようにして，当研究会ではマークアップの方針を完成させつつあり，今後はこれに則って大規模に人員を導入した作業を行うことができるものと考える．当然ながら，範例では扱いきれない，特殊なテキストの構造化に関する問題は残るものの，それらは絶対数として少数であり，大規模作業の妨げになるものではなく，個々対応すればよいものと思われる．まずはあまり細部に拘泥せずに，大部分のテキストに適応できる，大枠の範例の策定が第一に求められる．

謝辞

本研究には，JSPS 科研費 JP19H00516，一般財団法人仏教学術振興会及び大蔵経研究推進会議の助成を受けたものが含まれている．そして，SAT 大蔵経データベース研究会の関係者・関係機関の協力なしにはなしえないものであり，深く感謝申し上げる．

参考文献

[1] “大正新脩大藏經テキストデータベース”. https://21dzk.l.u-tokyo.ac.jp/SAT/, (参照 2020-11-09).

[2] “cbeta-org/xml-p5-2018”. https://github.com/cbeta-org/xml-p5-2018, (参照 2020-11-9).

[3] “P5: Guidelines for Electronic Text Encoding and Interchange, 12 Critical Apparatus.” https://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPDE ,（参照 2020-11-9）.