TEI P5 を利用した仏教用語集作成に関する諸問題

高橋 晃一
東京大学大学院 人文社会系研究科

近年，XMLは自由にタグを設定できるという柔軟性から，人文学の分野でもテキストの構造分析などへの応用が期待されている技術である．しかし，柔軟であるということはデータの共有の面で不都合な場合もあり，堅実なスタンダードが求められているのも事実である．この点に関して，TEI P5は非常に多くの種類のタグを提供しており，有望なテキスト・エンコーディング・スタンダードであることは疑う余地がない．しかしながら，実際にTEI P5のガイドラインに従って作業を行う場合に，問題に直面することもしばしばある．本稿では，XMLを用いて「仏教用語集」を作成する際に生じた，いくつかの具体的な問題とその解決策について報告する．

Some Issues on Encoding of the *Glossary of Buddhist Terminology* by Using TEI P5

Takahashi Koichi
Graduate School of Humanities and Sociology
University of Tokyo

Currently, XML is widely used in the field of Humanities, because this technology makes it possible to analyze the complex contents and structure of text flexibly by using tags which can be freely decided by users. However, this flexibility requires establishing a standard of text-encoding in order to exchange and share datum. To meet this need, *TEI P5: Guidelines for Electronic Text Encoding and Interchange* seems to be a compelling option, because it provides numerous stable elements to encode materials treated in the area of Humanities. While acknowledging the wide range of applicability of the TEI P5, in this short paper I will point out some concrete problems which are encountered when a trying to apply the TEI Dictionaries Module in the case of a complex glossary of Buddhist technical terms.

1．仏教用語とその現代語訳の必要性

「仏教用語」と一口に言っても，その意味するところは漠然としている．ある場合には「南無阿弥陀仏」のような漢字で表現される日本の仏教思想の用語を想起するであろう．しかし，ここで言う「仏教用語」とは，仏教の起源であるインドの古典語，すなわちサンスクリット語で著された原典の中で用いられている仏教術語を指している．この仏教用語は，他のインド古典やインド思想と比べても，極めて難解なものが多い．それはいくつかの理由が考えられる．

紀元前5世紀ごろ，インドにおいて成立した仏教は，その起源からインド在来の思想に対するアンチテーゼという性格を持っていたため，開祖ブッダの言葉にはそもそも難解なものも少なくなかった．さらに，数世紀もの間，難解なブッダの言葉を伝承する過程で，仏教徒たちは煩瑣な教理解釈を構築していくことになる．その結果，多くの難解な「仏教用語」が生まれることとなった．

仏教用語が難解な理由はこれだけではない．現存する仏典の多くはサンスクリット語というインドの古典言語で著されているが，ブッダ自身はサンスクリット語で彼の思想を語っていたわけではなく，彼の母語であるアルダマーガディ語という古代インドの一方言を用いていたと考えられている．ブッダの言葉は時代を経てサンスクリット語に置き換えられていったが，その間に本来のサンスクリット語には見られない語彙や語法が現れた．こうした事情も仏教用語を難解なものとする一因となったと考えられる．

このような難解な用語を含む仏教思想は中国やチベットなどの東アジア，スリランカ，ビルマ，タイなどの東南アジア諸国への伝播に伴い，それぞれの国の言語に数世紀に渡って翻訳された．さらに今日では欧米諸国でも仏教思想への関心が高まり，仏教文献を英語などの欧米諸語に翻訳する試みがなされている．

こうした状況の下で，文献学的視点から仏教思想を考察しようとする場合，サンスクリット語の難解な仏教用語をいかに現代語に翻訳するかが大きな課題となっている．日本を含めた漢

字文化圏においては，『西遊記』の三蔵法師のモデルとも言われる翻訳僧・玄奘（602-664）による漢訳語を符牒的に援用することが多い．しかし，そうした漢訳の仏教語そのものも，現代日本語としては意味をなさない場合や，本来の語義とはかけ離れた意味で日本語の中に取り込まれている場合も少なくない．一方，仏教用語の難解さは欧米諸言語への翻訳の際にも大きな壁となっており，サンスクリット語の原語を外来語としてそのまま取り入れている例もしばしば見受けられる．

積極的な情報発信が求められる現代において，最新の仏教学の研究成果を，広く一般の人々に公開していくことを目指すとき，従来の符牒的な漢訳語や，原語をそのまま用いる手法は必ずしも有効なものとは言えない．そのため，現代語としてわかり易く，また原語の意味を損なわない訳語を再考する必要がある．

2. 仏教文献研究と情報処理技術

仏教用語の現代語訳を考案するに当たり，原典資料を精査し，それに基づいて意味を考察する必要があることは言うまでもない．そのためには，ある術語に関する説明文や端的な用例などを原典から抽出し，データベース化する手法が有効であろう．

仏教文献は長い歴史の中で，サンスクリット原典から古典漢文，チベット語に翻訳されているほか，スリランカを中心とした東南アジア諸国ではパーリ語により仏典が伝承されてきた．このように，仏教文献の多くは古典期において既に複数の言語に翻訳されている．また，サンスクリット語原典に対して，インドで作成された注釈書のほか，各言語の翻訳文献に対して，各地域で著された注釈書も残っている．

このように，仏教文献のテキスト状況は非常に複雑で，これらの関係を踏まえたデータベースを作成することは多大な困難が予想される．そこで今回は，様々な状況に柔軟に対応することが期待される XML を用いて，複雑かつ多言語に関わるデータの整理を試みることとする．

XML を使用する際に最も大きな問題となるのは，タグの設定であろう．独自のタグを設定する場合，仏教学を研究する者の視点で利用しやすいタグを設定することが可能になる．しかし，将来的に他分野の XML データベースとの連携を考えた場合，データを共有する上でのデメリットも予想される．こうしたことを避けるために，今回は TEI（the Text Encoding Initiative Consortium）の最新のガイドライン TEI P5(1.8.0) [1]の 9.Dictionaries に従うこととする．

3. XML の利用と TEI P5 の有効性

XML を用いて語彙集を作成する場合，スタンダードとなるものを考えておく必要があることは言うまでもない．すでに述べたように，今回は TEI P5 をスタンダードとして採用しているが，この TEI P5 に準拠する有効性について，最新の研究成果が報告されているので，まず，その内容を簡単に紹介しておく．

2010 年 7 月にロンドンで開催された Digital Humanities 2010 において，Piotor. Bański 氏によって，ポーランド語のコーパスを XML を用いて作成する試み（the National Corpus of Polish, NKJP; http://nkjp.pl/）についての報告がなされた．同氏は，CLARIN(Common Language Resources and Technology Infrastructure)のショートガイド（2009 年 5 月）[3]で紹介された三つのテキスト・エンコーディング・スタンダード（ISO Process, TEI, XCES）を取り上げ，コーパスを作成する上でのそれぞれの特徴を，概略次のようにまとめている．

XCES について

2000 年代初期にはもっとも一般的なコーパス・エンコーディング・スタンダードではあったが，近年では影を潜めつつある．特に初期の XCES の特徴であった，形態統語論的データ構造の具体的な記述法は一般的で抽象的な記述法に置き換えられてしまい，コーパス製作者に負担を強いる結果となっている

ISO TC37 SC4 について

素性構造の記述法に関するスタンダード(ISO 24610-1)と辞書のエンコーディングに関するスタンダード(ISO 24613)を公表している．それらは将来性はあるかもしれないが，現時点では実用的なものではない．

TEI P5 について

メタデータと構造分析に関して勝れているだけでなく，ISO の提唱する素性構造記述法も満たしている．TEI P5 は安定しており，1350～1400 ページにわたる豊富な内容を提供している．現在これと比肩するものはない．

Bański 氏も指摘している通り，内容の充実度および安定性の面で TEI P5 は評価に値するものであり，またこれに準拠することで，他のエンコーディング・スタンダードとの互換性もある程度保障されると考えてよいであろう．少なくとも，具体的な内容や要素間の関係を簡潔に示すことができるタグを豊富に提供している点で，コーパスや辞書を制作する者にとって多大な利点を与えていることは確かであろう．

4．TEI P5 に準拠した用語集の作成例

それでは，実際に TEI P5, 9.Dictionaies に従って語彙集を作成する手法について考えてみる．TEI P5 では，辞書について，国語辞典のような「単一言語辞書」と英和辞書のような「多言語辞書」を念頭において解説されている．ところで，すでに述べたように，今回試みているのは仏教用語の現代語訳を考えるための基礎資料のデータベース作成である．このデータベースは基本的には「原典テキスト」とそれ対する「注釈文献」，および「諸言語訳」によって構成されている．その中核となる部分は，対象となる術語とそれに対するサンスクリット語原文による説明文や実用例となる．これはサンスクリット語の単語をサンスクリット語で説明する「単一言語辞書」とも言える．一方，このサンスクリット語原文に対して，数種類の古典訳および現代語訳の情報を付加することも今回の用語集作成のために必要なのだが，これはサンスクリット語の術語を，他言語で置き換えたり，説明したりするものであり，言い換えれば，「多言語辞書」の形態をとることになる．したがって，今回の語彙集は「単一言語辞書」と「多言語辞書」の両者の性格を持っていることになる．

ところで，そもそも辞書の各項目は「見出し語」と「解説文」でできている．TEI P5 では，この「見出し語」と「解説文」のひとまとまりを登録語彙に関する情報として<entry>タグで囲み，その子要素として，「見出し語」の基本語形の情報を<form>タグで，「解説文」を語の意味に関する情報として<sense>タグで示すことになっている．この三つの要素が辞書を作成する際の基本となる．語形に関してはさらに<form>要素の子要素として<orth>要素を置き，単語の正書法を示すことができる．また，意味情報に関しては，厳密な意味での「解説文」のほか，「語源解釈」や「用例」を記入することもあり得るが，当該のテキストが「解説文」あるいは「定義文」であること明示するために，<sense>要素の下に<def>要素を置き，その内容を規定することができる．例えば，サンスクリット語で「信心」を意味する śraddhā という語に関して，上の形式を当てはめると次のようになる．

```
<entry>
  <form>
    <orth>śraddhā</orth>
  </form>
  <sense>
    <def>śraddhā cetasaḥ prasādaḥ/</def>
  </sense>
</entry>
```

図１．単一言語辞書型（定義・語釈）

これは先に挙げた「単一言語辞書」の最も基本的なスタイルに相当する．ちなみに，「語源解釈」を記述する場合は，図２のように<def>タグの代わりに，<etym>タグを使用する．

```
<entry>
  <form>
    <orth>śraddhā</orth>
  </form>
  <sense>
    <etym>
      <!-- śraddhā の語源説明 -->
    </etym>
  </sense>
</entry>
```

図２．単一言語辞書型（語源説明）

また「用例」を示す場合は，書誌情報を含む引用を示すための<cit>タグと type 属性を組み合わせ，<cit type="example">とし，その子要素である<quote>要素として，「用例」に相当するテキストを記述することができる．（図３）

```
<entry>
  <form>
    <orth>śraddhā</orth>
  </form>
  <sense>
    <cit type=”example” >
      <quote>
        <!-- śraddhā の用例 -- >
      </quote>
    </cit >
  </sense>
</entry>
```

図３．単一言語辞書型（用例）

図１から３で示した<def>，<etym>，<cit type=”example”>は<sense>要素の子要素として並列することもできる．

さらにこれに対する古典訳の情報を追加する場合も，やはり<cit>タグを利用し，属性値 @type="translation" によって<quote>内の引用文が翻訳であることを示す．例えば，上記の図１にチベット語訳の情報を付加すると，次の図４のようになる．

```
<entry>
  <form>
    <orth> śraddhā </orth>
  </form>
  <sense>
    <def> śraddhā cetasaḥ prasādaḥ/ </def>
    <cit type=”translation” xml:lang=”bo”>
      <quote>dad pa ni sems dang ba'o//
    </quote>
      <bibl><!-- 書誌情報 --></bibl>
    </cit>
  </sense>
</entry>
```

図４．多言語辞書型

なお，上記の図２のように<cit>タグに言語を示すための属性@xml:langを与えることで，当該テキストの言語に関する情報を示すことができる．その際，言語の表記はISO 639-1に従う（[1]TEI P5, vi Languages and Character Sets 参照）．

５．定義文に出典情報を追加する

TEI P5 が扱っているのは一般的な辞書であり，そのため単語の説明文に相当する<def>要素としてのテキストに関する出典を明記する必要性が想定されていない．しかし，今回の試みでは，サンスクリットの原典から術語の定義文を抽出し，語義説明に充てているため，出典情報は必ず明示しなければならい．この場合，問題になるのは，TEI P5 では，<def>タグは定義文を直接格納するものであり，下位要素のグループを構成しないとされている点である．この点は，<form>タグがその子要素として<orth>要素などを取ることができるのとは性格を異にしている（[1]TEI P5, 9.3.3.1, Definitions 参照）．そのため，基本的に<def>要素は配下に書誌情報などを含むことはないので，このままでは出典の情報を記述することができない．

この問題を解消する手段として，例えば，そもそも<def>タグを使用せず，図３と同様に<cit>と<quote>を用いて定義文を記述し，<quote>タグ内のテキストを<term>タグと<gloss>タグによって分析するという方法も考えられる（図５参照）．

```
<entry>
  <form>
    <orth>śraddhā</orth>
  </form>
  <sense>
    <cit xml:lang=”sa”>
      <quote><term>śraddhā</term>
    <gloss>cetasaḥ prasādaḥ/</gloss></quote>
      <bibl><!-- 書誌情報省略 --></bibl>
    </cit >
  </sense>
</entry>
```

図５．出典情報を付した定義文（１）

<gloss>タグは，これによって囲まれた語句が他の語句に対する定義や解説であることを明示することができる．注釈の対象となる語句は<term>タグで示され，一般的には ID を付すことで，<gloss>要素との関連付けがなされる（[1]TEI P5: Appendix, <gloss>の項を参照）．

この方法でも辞書項目の定義を明示することができるが，文書全体の可読性を考慮すると，やはり<def>タグを用いて定義文を記述する方が望ましいように思われる．以下は TEI P5 で明確に規定されている方法ではないが，<def>要素に書誌情報を付すため，<cit>タグで囲むという方法を試みている．

```
<entry>
  <form>
    <orth> śraddhā </orth>
  </form>
  <sense>
    <cit xml:lang=”sa”>
      <def> śraddhā cetasaḥ prasādaḥ/ </def>
      <bibl><!-- 書誌情報省略 --></bibl>
    </cit>
  </sense>
</entry>
```

図６．出典情報を付した定義文（２）

<cit>タグで<def>要素を囲む記述法は，TEI P5 にも見られる（[1]TEI P5, 9.3.3.1, Definitions）．その全体を図７として引用する．

```
<entry>
  <form>
    <orth>rémoulade</orth>
    <pron>Remulad</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
    <gen>f</gen>
  </gramGrp>
  <cit type="translation" xml:lang="en">
    <quote>remoulade</quote>
    <quote>rémoulade</quote>
    <def>dressing containing mustard and
herbs</def>
  </cit>
</entry>
```

図７．**TEI P5, 9.3.3.1 Definitions** の第二例

一見して分かるように，これは出典情報を明示するためのものではない．TEI P5 の説明によれば，多言語辞書において翻訳語（例：<quote>remoulade</quote> と <quote>rémoulade</quote>）を説明文（例：<def>dressing containing mustard and herbs</def>）から区別する書式を例示するものに過ぎない．しかし，この形式を応用すると，先の図６のように<def>要素の出典情報を追加することができる．

また，<def>要素を<cit>の子要素とすることで，<cit>要素に属性@xml:lang を加えて，<def>要素の言語の種類を特定できることも利点としてあげられる．

６．古典的注釈を追加する

次に<def>タグで囲まれた原文に対する古典的な注釈とその書誌情報を記述する方法について考えてみたい．一般の辞書では定義文にさらに注釈を施すことはまれである．したがって，TEI P5 もこのようなケースの扱いには言及していない．例えば<note>タグなどを使って注記として扱うことも考えられるが，付加すべき情報が書誌情報を含んだテキストであることを考慮すると，<cit>タグと組み合わせて，次のように表記するのが良いように思われる．

```
<entry>
  <form>
    <orth> śraddhā </orth>
  </form>
  <sense>
    <cit xml:lang="sa">
      <def> śraddhā cetasaḥ prasādaḥ/ </def>
      <bibl><!-- 書誌情報省略 --></bibl>
      <note type="gloss">
        <cit xml:lang="sa">
          <quote><gloss>kleśopakleśakalu
    ṣitaṃ cetaḥ śraddhāyogāt
    prasādati</gloss> </quote>
            <bibl>
              <!-- 書誌情報省略-->
            </bibl>
          </cit>
      </note>
      </cit>
  </sense>
</entry>
```

図８．古典的注釈情報を付した定義文

まず，<note>要素が直前の<def>要素に対する注釈であることを明示するために，属性@type="gloss" を付加しておく．<cit>要素は古典的注釈文献に関する書誌情報を含めた情報の集合であり，子要素である<quote>タグ内の文章は，必然的に注釈を含んでいることになる．その当該個所は<gloss>タグを用いて明示することができる．

ところで，この場合の<gloss>要素は，先の<note>要素の属性@type="gloss"とはやや意味が異なる．図８の例では，<note>として引用した文章全体が対象語句の注釈に相当するため，すべて<gloss>タグで囲むことになり，要素としての<gloss>と属性としての@type="gloss" に全く相違がない状態になっている．しかし，このように引用した注釈文全体がすべて<gloss>要素に相当する場合ばかりとは限らない．例えば，この例では <def> 要素は “śraddhā”, “cetasaḥ”, “prasādaḥ” という三つの単語からできているが，注釈の仕方によっては一つずつ単語を取り出して，解説を加える場合もある．そのようなケースを想定し，<note>要素内の<quote>として引用される注釈文のテキストについて，注釈対象となる語句（例えば“śraddhā”）を<term>タグで示し，それに対する注釈文を<gloss>タグで囲み，ID などを付して整理し，さらに“cetasaḥ”, “prasādaḥ”などについても同様の処理をしておくことで，より繊細な要求に耐えうると考えられる（<gloss>タグの詳細な用法については，[1]TEI P5: Appendix, <gloss>の項を参照）．

８．同一ジャンルの文献を類例として追加する

以上で，語彙集の基本的な構造は出来上がるが，文献学的な視点からは，これに類例が列挙されていた方がより便利なものとなる．類例もやはり何らかの文献からの引用なので，<cit>タグを用い，属性@type="example"とし，<def>要素の親要素に当たる<cit>要素と並列させて，図９のように記述することができる．

```
<entry>
  <form>
    <orth> śraddhā </orth>
  </form>
  <sense>
    <cit xml:lang="sa">
      <def> śraddhā cetasaḥ prasādaḥ/ </def>
      <bibl><!-- 書誌情報省略 --></bibl>
    </cit>
    <cit xml:lang="sa" type="example">
      <quote>
        <!-- 類例のテキスト -->
      </quote>
      <bibl><!-- 書誌情報 --></bibl>
    </cit>
  </sense>
</entry>
```

図９．類例を付した定義文（１）

９．異なるジャンルの文献の情報を追加する

これまでは仏教文献の中でも単一のジャンルに関わる文献の整理法のみを考えてきたが，場合によっては，ジャンルが異なる文献でも，辞書の項目として必要な情報もありえる．仏教という宗教思想あるいは哲学を扱う文献の性格上，ジャンルが異なるということは，「意味」の違いとして捉えることも可能であろう．したがって，図１０のように，<sense>要素レベルで区別し，その子要素として<usg>タグを置き，それぞれの<sense>要素がどのジャンルに関わる「意味」情報を表すものかを明示する（<usg>タグの詳細については[1]TEI P5, Appendix, <usg>の項を参照）．

```
<entry>
  <form>
    <orth> śraddhā </orth>
  </form>
  <sense>
    <usg>abhidharma</usg>
    <cit xml:lang="sa">
      <def> śraddhā cetasaḥ prasādaḥ/ </def>
      <bibl><!-- 書誌情報省略 --></bibl>
    </cit>
  </sense>
  <sense>
    <usg>yogacara</usg>
    <cit xml:lang="sa">
      <def><!-- 省略 --></def>
    </cit>
  </sense>
</entry>
```

図１０．類例を付した定義文（２）

７．翻訳者名の記述法の問題

さて，複雑で難解な仏教用語集を，TEI P5 に準拠して XML データとして作成する際の，基本的な構造に関する考察は以上である．ところで，TEI P5 が提供する非常に豊富なタグは複雑な文献分析にも極めて有用であることは間違いないが，さらに翻訳者に関する情報を端的に明示できるタグ（例えば，<translator>など）を使用できるようになれば，より一層利用しやすいものとなるであろう．現時点での TEI P5 では翻訳者名は<editor role="translator">と記述することになっている．しかし，translator（翻訳者）と editor（編集者）では，その役割がそもそも異なっている．<editor>で代用できるのは compiler（辞書編者）までであろう．少なくとも translator は editor よりも author に近いものであり，<author>が単独のタグとして用意されているのであれば，<translator>あるいはそれに代わる固有のタグが用意されるべきではないだろうか（詳細は[1]TEI P5, Appendix, <editor>の項参照）．

１０．おわりに

「単一言語辞書」と「多言語辞書」の統合，定義文に関する出典の記述法，定義文に対するさらなる注釈の追記など，TEI P5 ではそもそも想定されていないケースをどのように処理するかという問題を考察してきた．結論としては，TEI P5 で規定されているタグを組み合わせることで，これらの問題は解決される．今回の試みでは，文献学的視点から文献の内部構造，および関連する外部の情報（翻訳文献・注釈文献など）との具体的な関係に着目して整理することにより，広く適用可能な構造を創り出すことができた．逆に言うと，TEI P5 の提供する極めて多種多様なタグが，いずれも具体性に富んでいるので，対象の在り様を実態に即して分析し，エンコーディングすることを可能にしているとも言える．今回提示した形態が唯一の解答であるとは言わないが，本稿筆者の考える文献の内部構造や関連する諸文献との関係を極めて明瞭に整理することができたことは，TEI P5 の有用性を実証することにもなろう．

謝辞：東京大学教授 Charles Muller 先生に，お忙しい中，英文要旨を校閲していただいたことに対して，感謝の意を表します．

参考文献

[1] Burnard, L. and Bauman, S.eds.: TEI P5: Guidelines for Electronic Text Encording and Interchange P5, Version 1.8.0 November 5th 2010. (http://www.tei-c.org/Guidelines/P5/ 2010 年 11 月 15 日閲覧).

[2] Bański, P. and Przepiórkowski, A: "TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation" presented at Digital Humanities 2010 King ' s College, London, UK, 2010/7/8/ (http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-616.html 2010 年 11 月 15 日閲覧)

[3] CLARIN: Standards for Text Encoding, 2009, (http://www.clarin.eu/files/standards-text-CLARIN-ShortGuide.pdf 2010 年 11 月 15 日閲覧)

この研究は，日本学術振興会・科学研究費補助金・基盤 A「仏教用語の『日英基準訳語集』構築に向けての総合的研究」（代表者・斉藤明）の成果の一部である．