Vol.2015-CH-105 No.5
2015/1/31

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

高橋洋成[†1]　永井正勝[†2]　和氣愛仁[†3]

近年のコンピュータとインターネットの発展に伴い，これまでコンピュータ上では転写テキストとして扱われることの多かった言語資料を，音声，動画，画像と一緒に保存・利用することが容易になった．著者らは古代エジプト神官文字文書，楔形文字粘土板資料，近代日本語文典資料などの文字資料を統一的に扱うための画像データベースと，World Wide Web 上のプラットフォームを構築している．さらに，本プラットフォームは次の 2 点を目指している．(1) 本プラットフォームの内部に格納された言語資料および研究データを，外部に保存し共有するために Text Encoding Initiative (TEI)を利用すること．(2)データの保存と共有をさらに促進させるために，言語資料と研究データに関する RDF トリプルを生成し，Linked Open Data (LOD)として提供すること．本稿はこの 2 点について，現在までの取り組みと具体例を報告する．

Construction of the Platform for Grammatological and Linguistic Studies using Images, TEI and LOD

YONA TAKAHASHI[†1]　MASAKATSU NAGAI[†2]
TOSHIHITO WAKI[†3]

Recent development of the internet and computer technologies makes it easier to compile various linguistic materials such as audios, videos and photos with their transliteration texts together. The authors are constructing the general-purpose image database and the platform on World Wide Web for the linguistic studies on hieratic texts of Ancient Egypt, cuneiform tablets of Ancient Orient, and Modern Japanese Grammar Textbooks. The platform aims on the following two points: (1) to use Text Encoding Initiative (TEI) for preservation and sharing of the internal data with its outer world, (2) to create sharable RDF triples from the data to provide them as Linked Open Data (LOD). The current snapshot of the developing platform will be discussed.

1. はじめに

近年のコンピュータとインターネットの発展・普及に伴い，コンピュータ上で言語資料を扱う方法も大きく変わりつつある．従来，録音・録画された「生」の調査記録と，それを聞き取って解釈を施した転写・翻訳テキストとは，別個のデータとして扱わざるをえなかった．前者はしばしばアナログデータであり，保存や再生に大きなコストがかかることから，研究者の間で共有されるデータとしては専ら後者の転写・翻訳テキストが用いられ，「生」の調査記録は「お蔵入り」になることも少なくなかった．

しかし，近年は音声や動画のデジタルデータ化と，保存デバイスの大容量化が進んだことにより，研究者の間で「生」の調査記録を公開・共有するための環境が整いつつある．さらに，たとえば ELAN のようにa，音声・動画に対し転写・翻訳テキストを字幕として表示するなど，「生」の調査記録と言語研究の成果とを結合させることのできるソフトウェアも公開されている．

著者らにとって「生」の調査記録に当たるものは，楔形文字粘土板資料，古代エジプト神官文字文書，近代日本語文典といった文字言語資料である．従来，これらの資料をコンピュータ上で検索可能にするには，個々の字形や文字の並べ方，あるいは改行の位置といった表面的な要素を捨象した転写テキストを作成せざるをえなかった．文字のコード化について言えば，古代エジプト神官文字は未だ Unicode に含まれていない．また，楔形文字は Unicode 化されたものの，時代や地域による字形の違いは無視されている[1]．さらに，近代日本語文典は挿絵に付された文字の並べ方に特徴があるにも関わらず，転写テキストでそのことを再現するのは難しい．しかし，このような表面的な要素もまた，人間が文字あるいは文字群をどのように認識し，言語として解釈するかという人間の認知の問題と深い関わりがあると考えられる．それゆえ，文字言語資料における表面的な要素をなるべく捨象することなく，かつ，語彙・文法をはじめとする幅広い言語研究にも耐えうるデジタルデータ化の方法が強く望まれている．

こうして，2012 年に高細度画像を用いた古代エジプト神官文字文書のデータベースとして開始したプロジェクトは，2013 年に文字言語資料のためのアノテーション付与型画像データベースおよび汎用プラットフォームとして発展し，2014 年にこのプラットフォーム上に近代日本語文典資料

†1 筑波大学
University of tsukuba
†2 筑波大学
University of tsukuba
†3 筑波大学
University of tsukuba

a https://tla.mpi.nl/tools/tla-tools/elan/

と楔形文字粘土板資料が追加された．本プラットフォームの目指す「汎用性」は，古代エジプト，古代メソポタミア，近代日本という広範な地域の，大きく性格の異なる文字言語資料を取り込むことにより，今後さらに精錬・発展していくことが期待される．それだけでなく，本プラットフォームは Text Encoding Initiative (TEI)および Linked Open Data (LOD)を利用し，文字言語資料を多くの研究者と効果的に共有していくことを目指している．では，本プラットフォームにおいてなぜ，どのように TEI と LOD が用いられるのか．本稿はこの点について現状を報告する．

2. プラットフォームの仕組み

2.1 内部構成

本プラットフォームの内部構成は[2][3]に詳しいが，ごく大まかに述べると，資料画像の処理を行う Zoomify，画像上の文字や言語の解釈をアノテーションとして管理する MySQL, これらを統合し Web ページとして出力する Drupal の 3 つのシステムから構成される．Web ページに表示された資料画像にポイントすると，その座標を含む一定の範囲がハイライトされ，その範囲に関連する文字情報・言語解釈情報が別ウィンドウに表示される．範囲はポリゴン座標で指定されているため，神官文字や楔形文字のように必ずしも矩形でなく，しばしば他の文字と重なり合うようなものも問題なく指定できる（図 1）．また，近代日本語文典のように挿絵に合わせて文字が斜めに並んでいるようなものも，1 つの文字にポイントすれば文字の並び全体がハイライトされ，それらがひとまとまりであることが容易に分かる．

図 1 矩形ではない神官文字
Figure 1 A non-rectangle Hieratic character

2.2 URI 設計

現在のところ，古代エジプト神官文字資料，近代日本語文典資料, 楔形文字粘土板資料の各プラットフォームは「共通の枠組みと構成であること」を意識し，wdb というサブドメインの下に，それぞれの名前空間を示すパスによって配置されている．

- Hieratic Database Project（図 2）
  https://wdb.jinsha.tsukuba.ac.jp/hdb/
- 近代日本語文典集成（図 3）
  https://wdb.jinsha.tsukuba.ac.jp/jgt/
- Cuneiform Tablets（公開予定）
  https://wdb.jinsha.tsukuba.ac.jp/xsux/

また, TEI および LOD において拡張語彙などを定義する際の名前空間を次のように設ける．

- 拡張語彙用の名前空間
  https://wdb.jinsha.tsukuba.ac.jp/vocab/

この URI 設計により，プラットフォームに共通する拡張語彙などを/vocab/に，各プラットフォームに必要な情報を各パスの下に，それぞれ置くことができる．

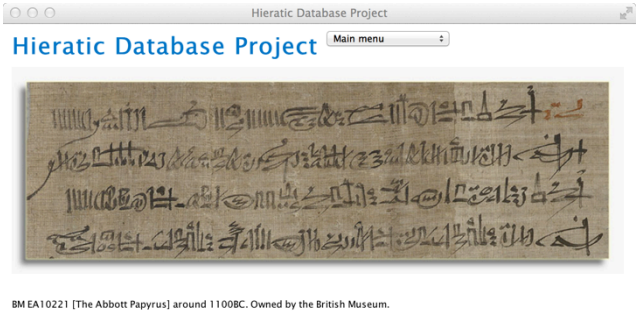

図 2 Hieratic Database
Figure 2 Hieratic Database

図 3 近代日本語文典集成
Figure 3 Modern Japanese Grammar Textbooks

3. TEI および LOD との関わり

3.1 なぜ TEI が必要か

本プラットフォームの内部では，資料画像，アノテーションの範囲座標，アノテーション情報が，それぞれのシステムに最適な形式で格納されている．一方で，資料画像とアノテーション情報を本プラットフォームの外部に保存し，特定のシステムに依存しない「文字言語資料自体」として研究者の間で共有・交換したい場合もある．このように，特定のシステムを想定せず，人文学資料自体を保存・交換・共有するためのガイドラインが TEI である．

本プラットフォームの内部システムでは，TEI 化されたデータを直接扱うことはない．しかし，本プラットフォームに格納された文字言語資料をパッケージングし，外部に保存するときに TEI が必要となる．そして TEI 化されたデータは，XML としての性質によって，XSLT などを用いて別のシステムに最適な形式に容易に変換しうる．言い換えれば，文字言語資料の外部への保存と共有は TEI に従って行い，各プラットフォームの内部ではそれぞれの処理に最適な形式で行うという設計を採用した．

TEI もまた，人文学資料のためのプラットフォームと見ることができよう．そうであれば，調査記録データやアノテーションデータが，それぞれのプラットフォームで共有され，行き交うモデルを考えることもできる（図 4）．

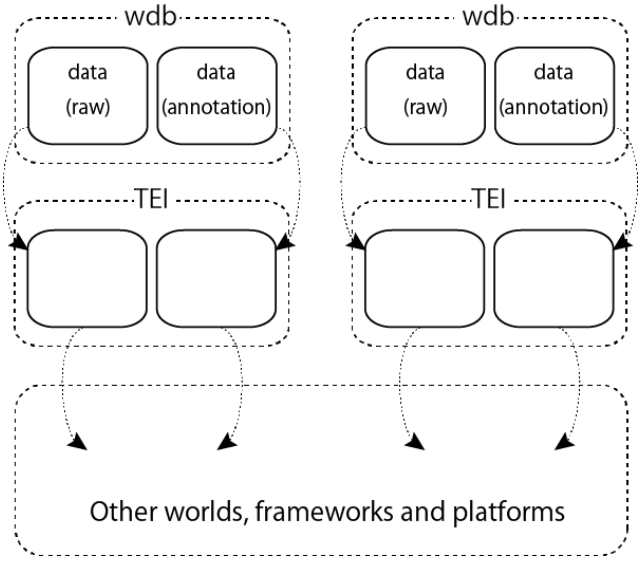

図 4 プラットフォーム間を行き来する言語データ
Figure 4 Linguistic data across platforms

3.2 どのように TEI 化するか

本プラットフォームにおける TEI 化の具体的なプロセスとしては，MySQL にアトミックな形で格納されている文字解釈・言語解釈データを，もう少し人間に読みやすい形に抽象化した TEI 文書を出力させれば良い．

ただし，TEI は人文学の多様な目的を考慮し，ある部分においてはデータ作成者に委ねるということが少なくない．たとえば，「語」であることを表す w 要素型は文法解釈を示す ana 属性を持つと定義されているがb, この属性値については「1 つ以上の xsd:anyURI 型」としか定められておらず，その URI が何を指すものかも定かでない．ガイドライン文書のサンプルコードでは，次に示すように，British National Corpus (BNC)で用いられている品詞タグをフラグメント識別子とし，それについての人間用の説明を同文書内の interp 要素に記述するというアイデアが記載されているc.

```
<s>
 <w ana="#AT0">The </w>
 <w ana="#NN1">victim</w>
 <w ana="#POS">'s</w>
 <w ana="#NN2">friends </w>
 ...
</s>
...
<interpGrp type="POS">
 <interp xml:id="AT0">Definite article</interp>
 <interp xml:id="NN1">Noun singular</interp>
 <interp xml:id="NN2">Noun plural</interp>
 <interp xml:id="POS">Genitive marker</interp>
 ...
</interpGrp>
```

とはいえ， 英語コーパスの作成を目的とする BNC の品詞タグは，本プラットフォームの目指す文字言語研究には若干不向きである．そこで，もしこのように URI を用いるのであれば，近年開発されている言語学用語の Web オントロジーとの対応を示すことで，言語学用語の意味を明示することができるだろう．以下は古代エジプト神官文字文書の一部を，General Ontology for Linguistic Description (GOLD)[4]の語彙を用いて記述した例である．なお現状，interp 要素に付す ID については，MySQL に格納されたフィールド名をそのまま出力している．

```
<s>
 <w lemma="pA" ana="#interp-lexical_category-5">
  <span>pA</span>
  <m ana="#interp-gender-1 #interp-number-1">
   pA</m>
  <note>the</note>
 </w>
 <w lemma="mr" ana="interp-lexical_category-18">
```

b http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.global.analytic.html
c http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILA

```
  <span>mr</span>
  <m ana="#interp-gender-1 #interp-number-1">
   mr</m>
  <note>pyramid</note>
 </w>
</s>
...
<interpGrp>
...
<interp xml:id="interp-lexical_category-5"
 corresp="http://purl.org/linguistics/gold/
  DefiniteArticle">Definite Article</interp>
<interp xml:id="interp-lexical_category-18"
 corresp="http://purl.org/linguistics/gold/
  CommonNoun">Common Noun</interp>
<interp xml:id="interp-gender-1"
 corresp="http://purl.org/linguistics/gold/
  MasculineGender">Masculine Gender</interp>
<interp xml:id="interp-number-1"
 corresp="http://purl.org/linguistics/gold/
  SingularNumber">Singular</interp>
</interpGrp>
```

言語学用語の語彙 URI については，このようになるべく共通化されたものを利用しつつ，必要に応じて拡張語彙を https://wdb.jinsha.tsukuba.ac.jp/vocab/ の下に定義する．

3.3 LOD への参加

前節で，言語学用語オントロジーの GOLD の語彙を TEI 文書の中に埋め込むことを検討した．さらに歩を進め，このように URI を語彙として利用するのであれば，文字言語資料自体を RDF として表現できないであろうか．そうすれば LOD を通じて，文字言語資料およびその研究成果の共有と発見をいっそう促進させることが可能になる．

まず考慮すべきことは，文字言語資料および解釈データにおける何を RDF 化するのか，ということである．著者らの主な関心は言語研究に置かれているため，現在の言語学用語オントロジーによって表現できる範囲，すなわち形態素，語，句，節，文といった言語構造的単位，および人称・性・数といった意味的特徴を，RDF グラフとして表現するよう試みた．また同時に，文字自体の仕組み，文字と語の対応関係といった文字論的情報を同じ RDF グラフの中に組み入れた．たとえば，図 5 は行（ex:line-1）の先頭から 5 文字（ex:c-1-1～ex:c-1-5）までの文字論的階層と，文（ex:s-1）の先頭から 2 語（ex:w-1～ex:w-2）という言語的階層のつながりを示したものである．

次に，どのようにして RDF を生成するかについて，3 通りの選択肢が考えられる．

A) TEI 文書に RDFa 属性を埋め込めるよう TEI スキーマを拡張する．

B) TEI 文書を XSLT 変換して RDF トリプルを生成する．

C) TEI 文書とは独立して RDF 文書を出力する．

(A)については単一の TEI 文書を出力すれば良いという利点があるものの，検討の結果，RDF トリプルを生成するための不自然な要素や情報の重複が避けられなかった．(B)については，TEI-P5 で明記されたスタンドオフ・マークアップ（リモート・マークアップ）と，RDF の構造との間に類似する点が多く[5]，変換自体は容易に行うことができる．ただ，RDF で言語学的な記述を行うための語彙[6]に比べ，RDF で文字論的な特徴を記述するための語彙が非常に限られているのが現状である．文字論的記述のための RDF 語彙については，別稿にて改めて論じたい．

なお，(A), (B), (C)の選択肢は必ずしも排他的ではなく，用途とコストに応じて複数を選ぶことは十分に可能である．本プラットフォームでは，(C)のように TEI 文書とは独立して RDF 文書を出力しつつ，同時に(b)のように TEI 文書と RDF 文書とを橋渡しする XSLT ファイルを用意した．

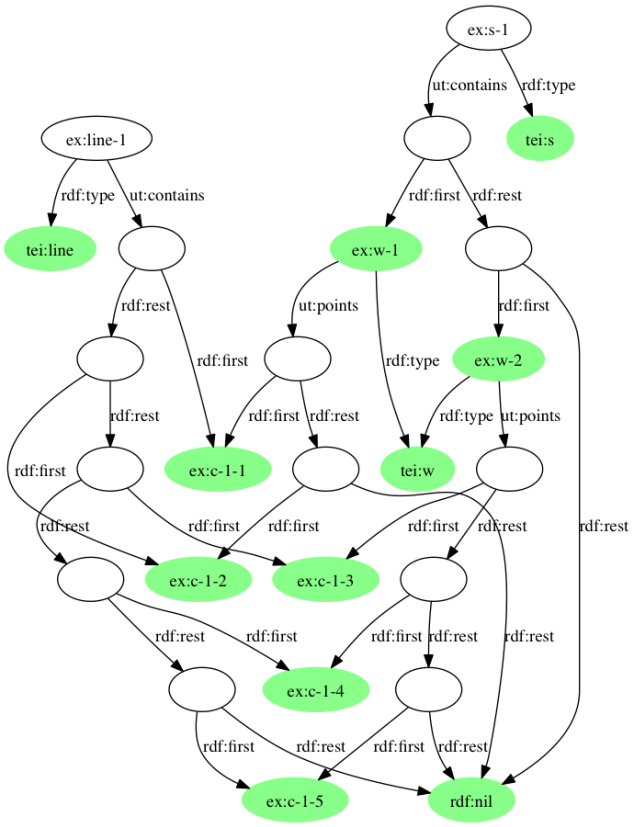

図 5 文字論的階層と言語的階層のつながり
Figure 5 Interaction between grammatological constituents and morphosyntactic structures

4. おわりに

本稿は，著者らが構築している文字言語資料のための画像データベースおよび汎用プラットフォームについて，内

部に格納された言語資料および研究データを TEI 文書として外部に保存し，言語資料自体を共有化する必要性と具体的な方法について論じた．また，言語学用語オントロジーをより柔軟に活用すべく，RDF トリプルを提供できるよう工夫し，LOD においてデータの発見が促進されることを目指した．

言語資料と研究データの共有化を押し進めることによって，本プラットフォームについても多方面からの研究者からのフィードバックを期待したい．

謝辞 本研究は科学研究費補助金「平成 26〜27 年度若手研究(B)：前 14 世紀の楔形文字文書，アマルナ書簡の言語記述のためのデジタルアーカイブ構築」代表：高橋洋成（課題番号：26870085），平成 24〜26 年度「基盤研究(C)：高細度画像と XML データを用いた古代エジプト語文書の言語記述アーカイブズの構築」代表：永井正勝（課題番号：24520452），および「平成 25〜27 年度基盤研究(C)：アノテーション付与型画像データベースシステムのための汎用プラットフォーム構築」代表：和氣愛仁（課題番号：25330395）の助成によるものである．数々のご助言や，貴重なデータを提供してくださった関係各位に，謹んで感謝の意を表する．

参考文献

1) 高橋洋成：アマルナ文書の電子化―文字研究・言語研究を目指して―，情報処理学会研究報告，人文科学とコンピュータ研究会報告 Vol.2013-CH-99, No.6, pp.1-7 (2013) ．
2) 永井正勝・和氣愛仁：古代エジプト神官文字写本を対象とした言語情報表示システムの試作，人文科学とコンピュータシンポジウム論文集, Vol.2012, pp.225-230 (2012).
3) 和氣愛仁：RDB と CMS を用いたアノテーション付与型画像データベースシステムの構築―データ構造とインターフェイスの標準化を目指して―，情報処理学会研究報告，人文科学とコンピュータ研究会報告, Vol.2013-CH-99, No.7, pp.1-8 (2013).
4) Farrar S. and Langendoen D. T.: A Linguistic Ontology for the Semantic Web, GLOT International, Vol.7, No.3, pp.97-100 (2003).
5) 高橋洋成：言語の多面性を織り込んだ言語資料のデジタルネットワーク，人文科学とコンピュータシンポジウム論文集，Vol.2013, pp.39-44 (2013).
6) Chiarcos C., Nordhoff S. and Hellmann S.: Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata, Springer (2012).