2007－CH－74 (2)
2007／5／25

漢讀：新しいテクスト・モデルに基づいた東洋学文献研究の支援ツール

ウィッテルン・クリスティアン、中楯はまな

京都大学人文科学研究所

この発表では、近代以前の中国語文献の研究を支援する為のツールを紹介する。こうしたツールを開発した動機は主に人文研で進行中の唐代ナレッジベースにあるが、いずれは一般化を目指している。マークアップの細かい手続きへの気がかりを必要とせずに、研究者がテクストでの研究対象に焦点をおくことができることが最大の目的だ。８つの観点からなるデジタル・テクストのモデルの実装に基づき、データ交換のためにTEI P5形式のXMLへの入出力ができる。

KanDoku: A tool for the study of premodern Chinese texts based on a new model of texts

This paper describes a partial implementation of a prototype of a tool for the study of premodern Chinese texts. The need for this tool arose, and the current work is conducted, in the context of the Tang Knowledgebase project at the Institute for Research in Humanities. The main purpose is to allow the researcher to focus on the interaction with those aspects of the text that are interesting to him, without the need to worry with the details of XML markup. It is based on an extended model of a text, that recognizes 8 facets of a text that needs to be implemented and supported in some ways by the software. It does allow import and export into generic TEI P5 XML.

1 はじめに

デジタル・テクストは書くことの起源以来知られている、従来のテクストとは根本的に異なる。　重要な違いを二つ挙げると、デジタル世界に入るためにはテクストの中の各文字は一つずつ解釈され、認識され、符号化されなければならない。　もう一つの違いは、デジタル・テクストは新聞や本のようにその場で読むことが出来ず、読む、統計を取る、検索する、形式を変換する等の使用のために、まずデジタルの形で保存されたものを再び読み出して、それから正しく解釈しなければならない。　普通はそれをソフトウエアで行うが、例えば電子辞書等のハードウエアによる解釈もある。　このような解釈ソフトの代表的な例はウエブ・ブラウザである。　正しい解釈が無ければデジタル・テクストは全く使えない。　そのため、デジタル・テクストのモデルとそのデジタル・テクストのモデルに基づいて機能するアプリケーションは同時に考える必要がある。

デジタル・テクストのモデルを考察する前にテクストの次元と側面を取り上げて、その中からモデルに採用する特徴を熟考する。

2 テクストの次元と側面

David Bolters (1991, [2]2001)と Jerome McGann (2001)を始め、1990 年代以降ハイパーテクストと情報時代のテクストを巡り、テクスト性(textuality)について、文学作品、特に詩、或いは編集学(Peter Schillingsburg 2006)を中心とした議論が盛んにされている。ここではこうした議論を踏まえて、テクストがもっているテクスト性の諸次元を列挙する試みである。全てのテクストはこうした次元を持っていると思うが、テクストの全体を見る時にはそれほど区別する必要はないので普段は意識していない。　デジタル媒体はこれにはっきり焦点を与え、それについての考察を押し進める。ある次元は漢籍ではもっとはっきり見えるか、視覚的な次元、意味的な次元と音声的な次元はほとんどのアルファベット言語ではそれほど明確に区分されないが、漢籍の場合では非常に重要な次元である。間テクスト性の次元も注意すべきである。この次元も中国の文化圏では非常にはっきりとした状態で現れる。また、例えば、テクストが現実に起こった事を情報として報告するかどうかといった別の次元もテクストのジャンルに依存する。　ここで取り上げたい８つの次元は以下の通り(図１参照)である。もちろん、互いに排反することではなくて、計算機上の実装のために焦点をそれぞれの次元に与える。

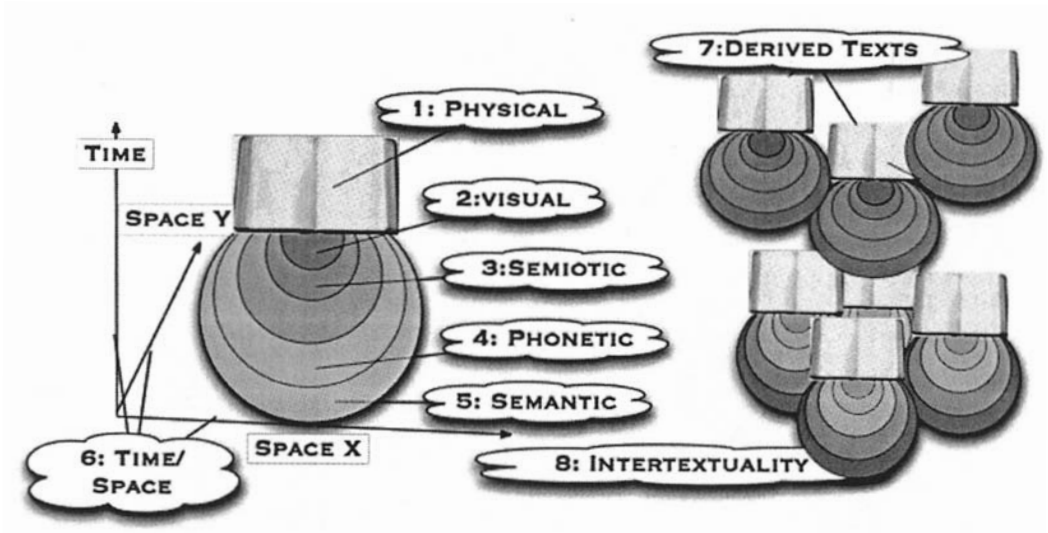

図 1

(1) 物理的な次元 (Physical dimension)

これは物理的に、物質的なモノとしての本のこと。この次元のデジタル化は全く不可能だが書誌的な記述ができる。記述はある具体的な、実際にあるモノの記述である。

(2) 視覚的な次元 (Visual dimension)

特定のテクストで使われた文字（正字、異体字又はその書体）、フォント、スタイル、ページレイアウトなどの視覚的な次元。

(3) 記号的な次元 (Semiotic dimension)

具体的に使われていた文字（異体字）から記号の次元を絞って抽象化した次元。デジタル・テクストの場合はこの次元は転写されて符号化の対象となる。

(4) 音声的な次元 (Phonetic dimension)

声を出して読まれるなら、そのテクストから読まれた次元。中国語の場合は反切、注音符号、またラテン文字への転写等、幾つかの記述記法がある。日本語の場合は仮名、或いはローマ字である。デジタル・テクストの場合はデジタル・録音も考えられる。

(5) 意味的な次元 (Semantic dimension)

これらがテクストのもっとも理解されている次元である。ほとんどの熟考がここから始まりここで終わっている。例えば、テクストで言及された出来事、地名、人名も含んでいるが、詩、小説、哲学的な論文、作文の内容的な扱いはこの次元に分類する。

(6) 時間的、空間的な次元 (Temporal and spatial dimension)

テクストは時間と空間において、特定のポイントで生まれる。そして、生まれて以来ずっと自分の人生を請け負う。これらの次元には粒度がもちろんあって、すなわち、それらはテクストのある一定のセクションだけに適用する可能性がある。また、時間次元は少なくとも以下の側面を含んでいる:

- テクスト（或いはテクストの一部）が作られた時点
- テクスト（或いはテクストの一部）に変更があった時点
- テクストの中で見られる時間的な側面（例えば歴史的な記録、ある出来事の報告など）
- 読者がテクストを読んでいる時点

(7) 派生テクスト (Derived texts)

ほかのテクストによって（影響を受けて）作られたテクストは《派生テクスト》('derived text')と名づける。例としては、あるテクストの注釈、翻訳、概要などが上げられる。仏典の異訳では少し複雑な例が幾つかある。『楞伽阿跋多羅寶經』（ーAD 443 年、求那跋陀羅（劉宋）訳）、『入楞伽經』（ーAD 530 年、菩提流支（北魏）訳）、『大乘入楞伽經』（AD 670~700?、實叉難陀（唐）訳）は全て『*Lañkāvatārasūtra』（成立は AD 400 年前後）の漢訳とされるが、恐らく原文もそれぞれの訳の間に変更があったとされる。間テクスト性との境界線がはっきりしていない場合があるだろう。

(8) 間テクスト性 (Intertextuality)

間テクスト性はここでは広い意味でテクストとテクストの間の関係を指している。例えば（部分的な）新しい組み合わせ、さりげない示唆、もじり、など。本テクストからの引用も本テクストでの引用も間テクスト性の次元として考える。

3 計算論的なモデルに向かって

デジタル方式で前述のような次元が実現するために、それらは電子媒体でモデル化されなければならない。いかなるデジタル媒体においても最も基本的なレベルでは、0と1しかあり得ないので、すべての情報は0と1の順序とパターンから読み出さなければならないことになる。他方では、媒体は非常に多目的に利用可能でフレキシブルになるが、音楽、写真、映画、テクストなど、本質的には違う形式の作品が、単なる0と1のパターンでほとんど同じ形式に変わる。このパターンの正しい解釈はデジタル情報の符号化と解読の根本的な課題である。

上記のように多方面で記述されるべきであるデジタル・テクストを符号化するためのモデルは階層を持ち、互いに依存関係を持つレベルに分けているモデルにならなければならない。例を挙げると（モデルA）、典型的で近代的な一冊の本は何枚かのページと装丁（表紙）からなる、一ページには何行か並んでいる、各行に言葉が並んでいる（実際は言葉が行、或いはページをまたぐこともある）、余白と句読点で分離された、最も小さい単位は全て文字として認識される。

図2：原文


```
<?xml version="1.0" encoding="UTF-8"?>
<div><lb n="7462-01"/>
 <p xml:id="zztj8-7462-p02d">初，<rm>懷光</rm>之解<dm>奉天</dm>圍也，
  <note place="inline">事見二百二十九卷<y n="建中-4">建中四年</y>。</note>上以其子
  <rm>璀</rm>為監察御史，<note place="inline">璀，七罪翻。監，<lb n="7462-02"/>
  古銜翻。</note>寵待甚厚。及<rm>懷光</rm>屯<dm>咸陽</dm>不進，
  <note place="inline">事見上卷<y n="興元-1">興元元年</y>。</note>
  <rm>璀</rm>密言於上曰：<q>臣父必負陛下，<lb n="7462-03"/>願早為之備。臣聞君、父一也；
   <note place="inline">人生在三，事之如一，謂君、父、師也。</note>
   但今日之勢，陛下 未能誅臣父，<lb n="7462-04"/>而臣父足以危陛下。陛下待臣厚，
   <app resp="章"><rdg wit="乙十六行本 乙十一行本 孔本">臣</rdg></app>
   <ym>胡</ym><note place="inline" type="ok">【<rm>章</rm>：
    乙十六行本<q>胡</q>上有<q>臣</q>字；乙十一行本同；
    <rm>孔</rm>本同。】</note>人性<lb n="7462-05"/>直，故不忍不言耳。</q>
  上驚曰：<q>知卿大臣愛子，當為朕委曲彌縫，而密奏之！</q>
  <note place="inline">為，于偽翻；下<lb n="7462-06"/>
  同。言璀當委曲彌縫，使君臣之間無隙，不當密奏其事。</note>對曰：
  <q>臣父非不愛臣，臣非不愛其父與宗族<lb n="7462-07"/>也；顧臣力竭，不能回耳。</q>
  上曰：[...]</p>
</div>
```


図3：モデルBの例


```
<page n="7462">
 <line n="7462-00">     資治通鑑卷第二百三十二 唐紀四十八    德宗興元元年(乙丑、七八五)          七四六二</line>
 <line n="7462-01">  初，懷光之解奉天圍也，事見二百二十九卷建中四年。上以其子璀為監察御史，(璀，七罪翻。監，</line>
 <line n="7462-02">古銜翻。)寵待甚厚。及懷光屯咸陽不進，(事見上卷興元元年。)璀密言於上曰：「臣父必負陛下，</line>
 <line n="7462-03">願早為之備。臣聞君、父一也；(人生在三，事之如一，謂君、父、師也。)但今日之勢，陛下 未能誅臣父，</line>
 <line n="7462-04">而臣父足以危陛下。陛下待臣厚，胡【章： 乙十六行本「胡」上有「臣」字；乙十一行本同；孔本同。】人性</line>
 <line n="7462-05">直，故不忍不言耳。」上驚曰：「知卿大臣愛子，當為朕委曲 彌縫，而密奏之！」(為，于偽翻；下</line>
 <line n="7462-06">同。言璀當委曲彌縫，使君臣之間無隙，不當密奏其事。)對曰：「臣父非不愛臣，臣非不愛其父與宗族</line>
 <line n="7462-07">也；顧臣力竭，不能回耳。」上曰：[...]</line>
</page>
```


図4：モデルAの例

このようなモデルは（２）に基づいて（３）からも影響を受けている。別の例（モデルB）では、文字から始まり、言葉、文、段落、節、章、本の順で意味としての単位を記述する。こういったモデルでは（５）と（３）に基づき（２）は除外される。

(A)と(B)のモデルの内、どちらが正しい、或いはどちらがテクストに最も適切なのか、という問いではなくて、目的によってどちらが目的に応じて最も必要とされる次元を円滑かつ効果的に符号化しているかという問いが必要だ。　例えばOCRソフトウエアのためのモデルでは恐らく(A)のようなモデルが直接視覚的な次元を扱うので効果的だろう。しかし、テクストの内容を主に扱う場合（大抵のテクスト処理）には（B）のようなモデルが有効だと思う。もちろん、それ以外のモデルも考えられるので、理論的或いは実践的な理由で特典を与えることができない。ただXMLの基本となる、テクスト処理で一番広く使われているモデルである「順序を持つコンテンツオブジェクトの階層としてのテクスト」(Ordered Hierarchy of Content Objects, [OHCO], Coombs et.al. 1987; Renear 2004)は少なくともテクスト・データの交換のために無視できないだろう。

現在進行中のプロジェクトの場合においては、基本的に（B）に基づいているが、漢籍を扱うために少し異なる要素の階層になる。この場合は文字から始めるが、その上にまず「句」を要素として置く。この「句」は文字の上に一番小さい意味を持つ単位である。句の上には同じような階層が有るが、今回はそれを包含階層ではなくて関連リストによって関連付けを行う。　上記で述べたほかの次元はこの句に関連をつけて、必要に応じて範囲（関連付けは句の全体では無くて一部だけである場合）を指定する。テクストとそれに直接反映させていない次元の関連は所謂スタンドオフ・マークアップ、つまり外からテクストの該当箇所を指す方法を採用する（図６参照）。

このモデルは根本的にはOHCOと互換性を持つので、XMLに基づいているTEIのタグセットに書き換えたデータ交換が可能である。

3.1 モデルから実装へ

どんなモデルを使っても、ユーザが必要とするテクストの次元をエンコードしても、実装が無ければユーザはそのモデルと接触ができない。実装によってある特定のタスクに特化してモデルの複雑さを隠す必要もある。　一方で、使用者のニーズ、問題意識、質問またはそれを計算論的に扱う方法は考慮しなければならないし、こういった条件は合わせて実装の方向性を導く。　最近のソフトウエア開発方針(Hale 2007)と合致させ、すぐに使えるプロトタイプをまず作って、それから使用者のフィードバックによって特化した機能を付け加える。

最初の要求は次の通り：

- テクストの閲覧、ブラウズ
- テクストの検索
- テクストの違うバージョンを見る
- 知らない言葉を調べる
- テクストの内容を分析する
- テクストに興味が有る要素をマークする
- テクストに注釈をつける
- テクストを翻訳する

この内の一部の要求が上記の「テクストの次元」に一致する。例えば「テクストの違うバージョンを見る」は上記の（1）の内の「書誌学的な記述」と（2）「視覚的な次元」とほぼ対応するし、「興味が有る要素をマークする」は「意味的な次元」と一致する。その他の項目は、このテクストの基本モデルと直接関係がなく、テクストを読む、理解する、分析する、注釈を付ける、翻訳するなどの研究目標からなる。

それ以外のユーザからの要求は辞書、書誌目録、人名、地名、地図など良く使われているツールの統合機能である。

4 カンドク（漢讀、KanDoku）

現在開発中の実装プロトタイプは「漢讀」（漢籍の読書支援ツール）と名付けた。この実装は商用データベース開発システム「ファイルメーカー」上で実現した。長期的にはクライアント・サーバー・アーキテクチャを想像しているが、モデルとその実装で直接実験できるため、現時点ではスタンドアロン型になっている。　上記の要求を実装する最中だが、それ以外の特徴としては上記のテクスト・モデル上で直接動作できる点と TEI P5 の XML からの読み込みと書き出しが可能な点である。この二点について少し詳しく説明したい。

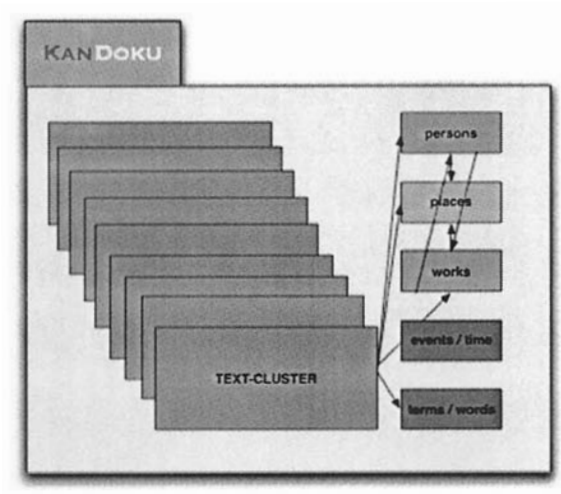

図５：KanDoku　概要図

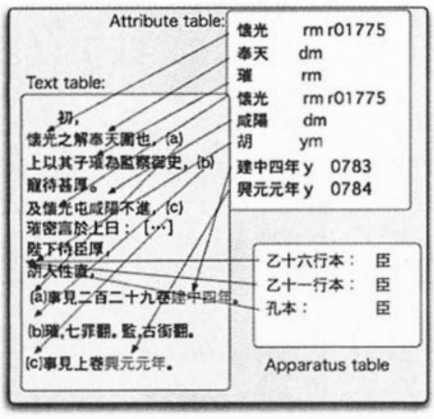

図６：KanDoku　テクスト・属性リンク

4.1 XML からの読み込みと書き出し

漢讀の XML 形式は通常の TEI スキーマよりもっと厳密なスキーマを必要とする。TEI では用途に従って TEI のスキーマをカスタマイズできるが、ここではその機能を使って、漢讀用のスキーマを作成した。漢讀で扱うのは TEI のサブセットである。漢讀用の XML ファイルは TEI のスキーマでも妥当だが、純粋な TEI は漢讀スキーマでは妥当ではない可能性もある。一番重要な制約はネスティングの制約である。テクスト中のネスティングはだいたいのケースでは段落の下で２レベルまでとなっており、フレイズ・レベルの一部のタグも扱っていない。読み込み時にはテクストを幾つかのテーブルに分けるが、多くの場合は上記のテクストの次元と対応する。例えば、一つのテーブルは正規化されているテクストを記録するが（3、

記号的な次元）、それに対しての違うバージョンが持っている文字のバリアントは別の、関連付けられたテーブルで記録する。このバリアントの内には実際のテクストが持っているバリアントもあるが、正規化の結果としてテクスト・テーブルで記録されたテクストと違う文字になる可能性もある。その場合は原文に合ったバリアントをこのテーブルで記録する。

4.2 テクスト・モデルの上で直接動作

複数のテクスト要素がマークされたら、XML テクストは複雑な構造で読み辛くなる。こういったテクストをそのまま編集するのは困難であり間違いが起こりやすい。その上に人名、地名、書名、などの「事実」の次元もマークしたり確認したり必要もある。今回の漢讀はこれらの問題を解決する試みでもある。　作業別に専用画面を作成して、その作業をできる限りインターフェースで支援して簡単に出来るようにする。　既製品のデータベース開発環境を使用するメリットの一つは、使用者自身でも比較的簡単に新しい画面を作成できることである。　それによっては漢讀の使用者が（もしくは人文科学各分野の研究者が）提供されたアプリケーションをそのまま使う「使用者(end user)」から、自分で必要に応じて新しいデータベースまで作成できる「作成者(end-maker)」になることが望ましい（cf. McCarty 2005, p15 et passim）。

この数ヶ月間、漢讀の開発バージョンは唐代ナレッジベースのマークアップ作業のみならず、漢訳仏典、中国中世の文学選集、法制史の資料等にも用いられた。

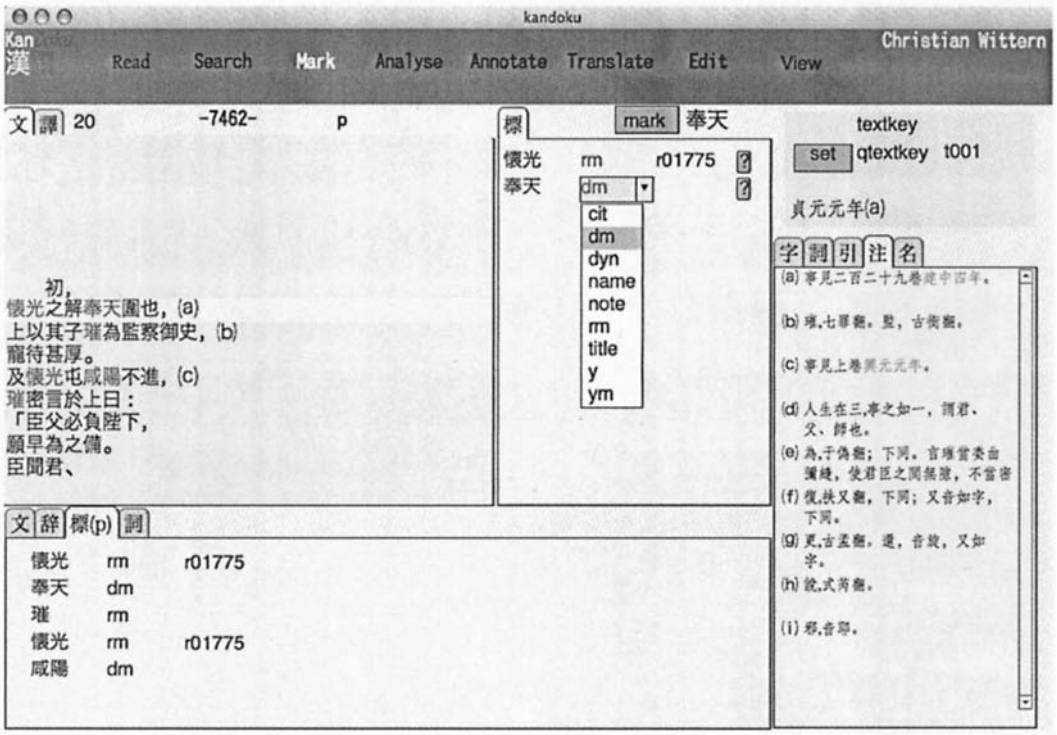

図７：漢讀の「マーク」画面

5 まとめ

概念上でデジタル・テクストをモデル化する為のモデルを提案した。このモデルはデジタル・テクストに対して、今まで統合されていない８つの次元をまとめて扱う試みである。この概念モデルを基にユーザのニーズを考慮して「漢讀」という実装のプロトタイプを開発した。異なるジャンルと異なる時代のテクストは実験的に漢讀に導入されながら、データベースのスキーマを修正し続けた。　問題点としては複数のテクストから大量テクスト・データベースへのスケーラビリティ問題が明らかになったことである。この点は今度の課題としてウェブ上のクライアント・サーバー・アーキテクチャを実装することで解決出来るだろう。

6 参考文献

David J. Bolter, *Writing Space: Computers, Hypertext, and the Remediation of Print*, Hillsdale, N.J (Lawrence Erlbaum Assoc), ^{2}2001 [1991]

Coombs, J. H. et al. "Markup Systems and the Future of Scholarly Text Processing", in Communications of the ACM, p.933-947, 1987. URL: http://xml.coverpages.org/coombs.html

Peter Hale, PhD Research User Driven Programming http://www.cems.uwe.ac.uk/~phale/

Willard McCarty, *Humanities Computing*, Houndmills (Palgrave McMillan), 2005

Jerome McGann, *radiant textuality. literature after the world wide web*, Houndmills (Palgrave McMillan), 2001

Allen H. Renear, Text Encoding, in: *A Companion to Digital Humanities*, eds. Schreibman et al. (Blackwell Publishing), 2004, p218-240.

Peter Schillingsburg, *From Gutenberg to Google*, Cambridge (Cambridge University Press), 2006

C.M. Sperberg-McQueen and Lou Burnard (eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Oxford et al., 2007 URL: http://www.tei-c.org/release/doc/tei-p5-doc/html/

Christian Wittern "The Text in the Age of Digital Reproduction", *The Role of Buddhism in the 21st Century. Proceedings of the Fourth Chung-Hwa International Conference on Buddhism*, p.389-414, Taipei 2005.