

Old Tibetan Documents Online (OTDO) の取り組みと課題

松田 訓典

星 泉

東京外国語大学アジア・アフリカ言語文化研究所

Old Tibetan Documents Online (OTDO, <http://www.a.tufts.ac.jp/>) は、およそ七世紀から十二世紀のものと推定される敦煌写本、碑文、その他関連する古チベット語諸資料のコーパスを作成・公開しているプロジェクトである。その収集・校訂作業はここ数年、比較的少人数ではあるものの、国内外の研究者によって、wiki をベースとしたオンライン作業という形で進められていたが、そこにはさまざまな課題が見られた。

本稿では、本プロジェクトの wiki ベースのシステムの紹介とそこで問題とされた点、それに対して今回新たに取り組んでいる新インターフェースの構築作業の進展について報告する。

Work of Old Tibetan Documents Online (OTDO)

Kuninori Matsuda

Izumi Hoshi

Institute for Languages and Cultures of Asia and Africa
Tokyo University of Foreign Studies

Old Tibetan Documents Online (OTDO) is a corpus of selected Old Tibetan texts (7th to 12th centuries) comprising Dunhuang manuscripts, inscriptions, and related materials. Formerly, the wiki-based system was used to collaboratively prepare the texts and related information. However, as the corpus expanded, certain problems or unsatisfactory points in the wiki-based system were observed. To alleviate these problems, we are developing another collaborative editing system.

In this presentation, we report the outline of our project, problems we encountered with the wiki-based system, and merits of the new system and areas for improvement.

1. はじめに

Old Tibetan Documents Online (OTDO, <http://www.a.tufts.ac.jp/>) は、およそ七世紀から十二世紀のものと推定される敦煌写本、碑文、その他関連する諸資料を、アルファベット転写・校訂し、古チベット語文書のコーパスとして作成・公開しているプロジェクトである。

そのコーパスの収集・校訂作業はここ数年、比較的少人数ではあるものの、国内外の研究者によって、wiki をベースとしたオンライン作業という形で進められていたが、同時に、対象となるテキストの増加等の要因もあって、作業工程上いくつかの不満点・改善されることが望ましい点が見出されるようになってきていた。

そこでそれらを解消すべく、昨年度より新たなインターフェースの構築を進めている。

本稿では、本プロジェクトの共同編集作業が開始されて以来の wiki ベースのシステムの紹介とそこで問題とされた点、それに対して今回新たに取り組んでいる新インターフェースの構築作業について報告する。¹

¹ 本プロジェクトの概要およびその基本的方針に関してはすでに拙稿[1]で報告したので、併せて参照されたい。

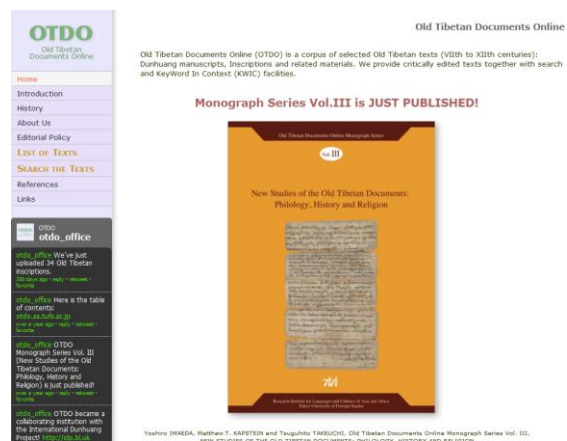


図 1: OTDO ウェブサイト

2. OTDO の歴史

本プロジェクトは、1970 年代にフランス国立科学センター (CNRS) の今枝由郎氏によって開始された古チベット語文書の集成を端緒としている。

1980 年代になり、当時東京外国語大学アジア・アフリカ言語文化研究所 (AA 研) が保有していた大型計算機のチベット語入力システムを利用して敦煌文書のデータを入力されることになった。その成果は 1990 年、文献[2]として出版されている。さらにその仕事は 1990 年代後

半になって AA 研の星泉に託され、同計算機を使った仕事を継続しつつ、2001 年には[3]が出版された。

その後、時代の変遷に伴い、大型計算機のデータをすべてテキストデータに変換し、それをもとに試行錯誤を繰り返しながら、web 上でも公開するようになった。当初はそのテキストをそのまま掲載するという形のものであったが、2003 年に最初の検索システムもつくられた。

これが OTDO の始まりであり、同時に、2001 年から 2006 年まで AA 研に置かれた COE プロジェクト GICAS (「アジア書字コーパスに基づく文字情報学の創成」, Grammatological Informatics based on Corpora of Asian Scripts) のサブプロジェクトとしても位置づけられていた。

2004 年ごろになり、wiki への編集作業の移行という、OTDO にとっての一つの大きな転機があった。これを契機にオンラインでの共同編集が開始され、現行システムの基礎となる新しい検索システムも作られた。同時にその成果は AA 研から OTDO モノグラフシリーズとして出版が継続されている(文献[4-6])。

また、近年 International Dunhuang Project (IDP, <http://idp.bl.uk>) との連携も進められており、その期待される役割を大きく広がっている。

3. 古チベット学における役割

ここであらためて述べるまでもなく、コーパスの作成は当該分野の研究にとってきわめて重要なものである。

特に本プロジェクトが扱っている碑文等に見られる古チベット語に関しては、語彙の面でも文法の面でも未だ十分に解明されていない部分が多く、非常に難解である。このようなテキストに関するコーパスを作成・公開することは、古チベット学のみならず、仏教学などの隣接する諸分野にとってもきわめて重要なものであるといえる。

チベット学の分野においても、研究上コンピュータ利用が不可欠になっているが、大部分の研究者はそれぞれ自分たちで作成したコーパスを限られた範囲の研究者たちで共有する形をとっているものとみられる。結果として、現状では、電子データとしてのコーパスの公開・共有を行っているプロジェクトは多いとは決して言えるものではない。

近年では、いずれもアメリカに本部を置く TBRC (Tibetan Buddhist Resource Center, <http://www.tbrc.org/>) や ACIP (Asian Classics Input Project, <http://www.asianclassics.org/>) といったプロジェクトがかなり大規模なテキストあるいは版本等の画像を電子化し、公開もしくは販売しているが、全文検索等の機能は提供していない

ようである。そういった点も含めて研究者ベースでデータを作成し、しかも完全に自由に共有できる形で公開しているという点も含めて、本プロジェクトの果たす意義は大きいと考えられる。

4. 従来の方式

ここで、従来行われていた wiki ベースの編集とその公開体制について紹介しよう。¹

はじめに、従来編集用の wiki と公開用のページは別に用意され、ある時点までの編集済みデータをまとめて公開用ページに反映させるという形をとっていた。

その中でまず公開用ページについて簡単に触れよう。公開用ページは主として、テキストのリストを表示するページ、テキストを検索表示するページ、リファレンスを表示するページからなる。

個々のテキストは図 2 のように、上部に資料の情報、下部に転写テキストが表示され、通常表示とともとの資料の物理的行単位での表示との切り替えが可能となっている。またさらに、このテキスト (ITJ_0730) の場合、International Dunhuang Project (国際敦煌プロジェクト, IDP) によって写本の画像が公開されているため、“photo (IDP)”という形で該当ページへのリンクが設けられている。なお IDP との連携については後述する。



図 2: 従来のテキスト表示ページ

テキスト検索(図 3)では、検索対象となる全テキストの一覧がチェックボックスとともに表示され、すべてあるいは一部のテキストを対象とした、前方一致等の検索や OR 検索が可能となっている。

その検索結果は、KWIC 形式で表示され(図 4: 一例として“bod yul”(チベット)の検索結

¹ 本システムは有限会社ツクラカンの千葉文博氏によるものである。

果)、当然各テキストへのリンクを含んでいる。また当該キーワードの前後の文脈でのチベット語順ソート機能ももつ。

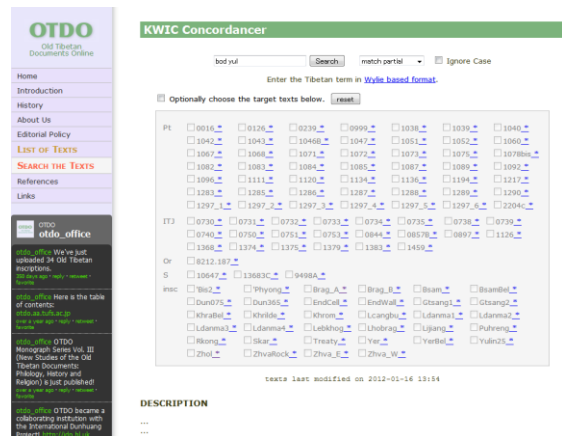


図 3: 従来のテキスト検索画面



図 4: 従来のテキスト検索結果

従来の検索公開ページの概要はおおよそこのようなものである。次に編集用の wiki について紹介しよう。

編集用の wiki に関してもおおよそ三種からなる。(a) 転写テキスト・校訂 (図 5), (b) テキストのカタログ情報 (図 6), (c) 参考文献 (図 7) である。

(a) に関しては上部にテキストに関する簡単な説明、テキスト全体に関わる編集者のコメント、行単位のテキストとその校訂に関するコメントからなる。

次に (b) に関しては、最上部に「仕様」としてカタログ情報入力に関する項目わけとその概要や入力の指針が示され、以下、テキストごとにそれぞれの内容が列記される形となっており、文献ごとに別ページとはなっていない。項目としては、文献番号・内容の簡易紹介・媒体の形状・テキストの状態・関連文献・IDP 画像のリンク等といったものが、元となる資料の種類に応じて定められている。

最後に (c) に関してもほぼ (b) と同様の構



図 5: wiki (a) 転写テキスト・校訂



図 6: wiki (b) テキストのカタログ情報



図 7: wiki (c) 参考文献

成をとっている。上部に「テンプレート」として、著者・論文名などといった書誌項目が定められ、それに応じた項目を、以下に列挙する形となっており、同じく参考文献ごとに別ページとはなっていない。

ここで編集されたテキストや参考文献のデータは、ある時点で公開用サイトにインポートされるという形をとっている。

5. 従来の方式の問題点と刷新へ

この方式には、問題とはいえないまでもいくつかの不具合・不足点が存在する。実際、編集作業を行う研究者たちからも以下のような意見が挙げられている。ここではおおそ主にシステム自体に起因する問題と、編集や検索といった機能面での問題にわけて紹介しよう。

〈システム面〉

- a1) テキストの複数人編集・工程管理
- a2) 参考文献等の編集に関してどこにどのようにかがわかりづらいことへの対処
- a3) 参考文献と各テキスト間でのクロスリファレンス
- a4) 編集作業中に検索ができないことへの対応
- a5) フォーマットの変遷による不統一の解消

〈機能面〉

- b1) 時代・ジャンル等の検索条件の絞り込み
- b2) IDP へのリンクのサムネイル化
- b3) 行をまたぐ検索
- b4) チベット語転写の n/ng, s/sh, z/zh の区別
- b5) 編集項目の保持（従来は公開用ページに一括インポートされる）

必ずしもシステム面・機能面と明確に区分できるものばかりではないが、当初、掲げられていた問題点はおおよそ以上のようなものであった。

既存システムは、wiki 上での編集とはいえ、基本的には紙ベースの作業と同様の入力内容をもっており、「サイト開始時より入力が続いているが、どの行にどの情報をどのように入力すべきか、いまだに迷うときがある。これから新編集者が参入するが、これを機にもう少し簡単な入力方法を考えたい」という意見も出ていた。

こういった点でシステム側からの補助を期待できる部分が多々あることが想定されるので、システムの刷新への取り組みが 2011 年度行われることとなった。

システムの刷新にあたって、まず問題となったのは、wiki システムを維持するかどうかであった。

wiki を維持する場合、適宜プラグイン等を準備し、フォーム等を設置し、しかるべき形でページを分割するという作業が必要になると思われる。

一方、新しくシステムを構築する場合、しかるべきインターフェースを作成し、データをインポートするという形になる。

wiki を維持する利点は、基本的にデータの移行等の手間がない、という点であろう。逆に、欠点としては、そもそも基本的にべた書きであることから、何らかの構造化を施そうとした場合にやはりある程度のデータ処理が必要になる点である。その際、既存の方式からの変化がユーザにとって混乱のもととなるかもしれない。また要望の一つにあった、編集作業中の検索を実現するためにも一工夫が必要となる。

一方、後者のメリットはなんととってもかなり思い切ったデータフォーマットの変更が可能点であり、クロスリファレンス等に関しても最初から折込済みの対応をとることができる。また参考文献編集等のインターフェースに関しては、筆者は別のところで作成した経緯があり、これを部分的に流用することも考えられた。

こうした点から、完全に必然であるとはいえないが、新しいインターフェースの構築を選ぶこととなった。¹ 既存 wiki から似て非なるものではなく、そこからの完全な脱却により、システム移行時の不要な混乱も避けるという面も考えられる。

6. 新方式の取り組み（1）

本節ではすでに実装済みの新方式の基本的な機能と、既存の問題に対する対処法を説明する。まず、従来の方式との最大の違いは、検索用ページと編集用ページの統合である。（図 8）

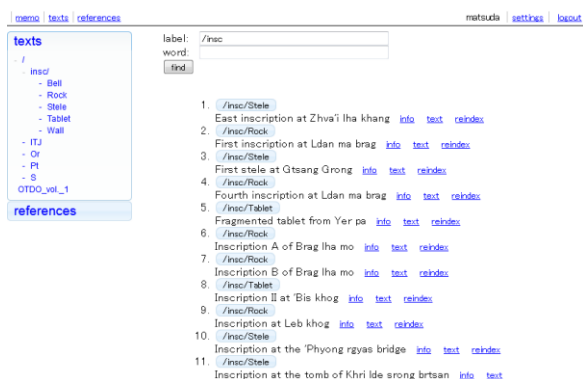


図 8: テキスト検索画面

このインターフェースの構成を簡単に説明すると、左サイドバーにテキストおよび参考文献のアコーディオン型メニューとなっており、一

¹ この点に関しては、近接分野ですでに国際共同編集プロジェクトを行っている永崎研宣氏からの助言もいただいた。ここに記して感謝の意を表したい。なお、同プロジェクトに関しては文献[7]参照。

方、右側には、主要な表示枠として、さまざまな情報が表示されることになる。

たとえばテキスト関係のものを表示する場合には、サイドバーには簡易ラベルメニュー（後述）が表示され、右側には検索結果等が表示される。

一般利用者はそのまま検索等を行い目的の情報を取得し、編集者サイドは各々のアカウントでログインすることで、同一インターフェース上から編集作業を行う形となっている。

この方式を採用することによって、まず a4) 編集作業中に検索ができないという問題は解消される。

■簡易ラベル機能

また本システムは簡易的なものであるが階層つきラベル管理機能を提供している。これによって、たとえば“/insc/Stele”といった場合、大分類としては碑文であり、そのうちの **Stele inscriptions** に含まれるものであることがわかる。この例はむしろ目次的な機能をもつものであるが、ラベル自体は編集者が自由に追加・編集でき、検索条件として指定できるものであるため、本機能を組み合わせることで、b1) 時代・ジャンル等の検索条件の絞り込みを行うこともできる。また現時点では二色であるが、随時色指定できるようにすることでわかりやすくすることもできるかもしれない。

また、特定管理者のみしか変更できないラベルとして、“入力中”“編集中”“校正中”“未公開”といった内容のラベルを準備することで、a1) テキストの複数人編集・工程管理、および、b5) 編集中項目の保持といった問題にも対処できる。ただし、特に前者に関しては専用の管理ページも設ける方がよいかもしれない。

■カタログ情報・参考文献情報

この点に関する問題を再掲すると以下の三つである

a2) 参考文献等の編集に関してどこにどのようにかがわかりづらいことへの対処

a3) 参考文献と各テキスト間でのクロスリファレンス

a5) フォーマットの変遷による不統一の解消

これらの問題は、基本的には、適正な入力フォーマットを用意し、適切にデータを移行すればおおそ解決する問題であると考えられる。そこで用意したのが、図 9 のインターフェースである。

この画面構成を簡単に説明すると左側上部がすでに説明した簡易ラベルメニュー、左側下部に新たに参考文献のラベル状ボタンリスト（各文献の詳細ツールチップあり）が現れ、右側に編集対象となる文献の情報が表示されている。ここで a3) クロスリファレンスの問題に関係す

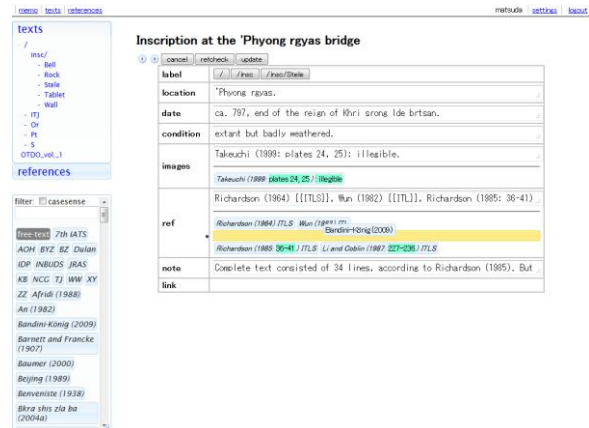


図 9: カタログ情報編集画面

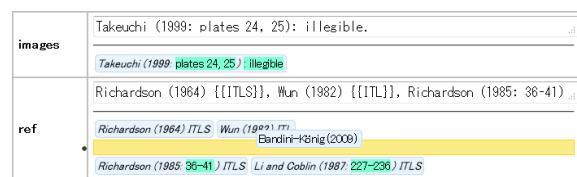


図 10: クロスリファレンス部分の拡大図

るのは、images（関連画像）および ref（参考文献）であるが、当該箇所を拡大したものが図 10 である。

現在はデータ移行のチェックのために、それぞれが、既存のテキストによる入力（上部）とそれをパースした結果としての新しいラベル状表示法（下部）という二段構成になっているが、チェックが終わり次第それぞれ上段は削除する予定である。

正しく参考文献リストに掲載されているものに関しては、図のように青系のラベル状ボタンとなり、一方登録されていないもの、あるいはその他自由記述テキストの場合には、グレー系のラベル状ボタンで表示するようになっている。

編集時は先の左下にあるボタン上リストからこの欄にアイテムをドラッグアンドドロップすることで、新しい項目を追加・並び替えを行い（図 9, 10 はその操作中）、個々の情報（個別のページ数や補足情報）を追加・編集する場合は、当該ラベル状ボタンを右クリックしてさらに編集画面を出す方式としている。

a2) に関してはフォーム形式にすることで、一定の解決を行い、必要に応じて、ツールチップ等によるヘルプの充実を図りたいと考えている。

既存データの整合性にかかわる a3), a5) に関しては、上述のように、基本的なパース処理を行い、しかるのち、不明瞭な個所を視覚的に表示することで、修正の機会を提供する、という方針をとる。

参考文献に関しても、基本的には同様の方針をとっている。



図 11: 参考文献検索の例

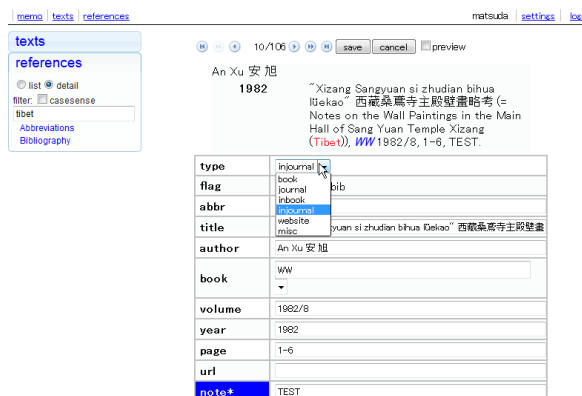


図 12: 参考文献編集の例

図 11 は検索画面，図 12 は個別の文献の編集画面である。

図 11 では，クロスリファレンス対象となる略号表記された論文収録雑誌名などを適宜ツールチップとして表示している。

図 12 では，参考文献の種別として雑誌・単行本・ウェブサイト等の種別を設け，各々に応じて動的にフォームを変えたり，収録雑誌等には自動補完機能を設けたりすることで入力の手を軽減する。またどこが今回編集した項目か，それがどのように反映されるか，等も確認できるようにしている。

■テキスト編集

テキストの編集に関しては，基本的な編集画面として図 13 のような形としている。

これに関しては，行単位でのフォームの導入，および，従来のテキスト単位での一括インポート機能を用意しているが，とりたてて目新しい要素といえるほどのものではない。あえて挙げるとすれば，従来手入力されていた校訂用の表記（異読や訂正など）の入力を補助する機構を用意する程度である。a5) フォーマットの変遷による不統一の解消に関しても，先述の基本的なパス→視覚化→修正の促進，という流れを意

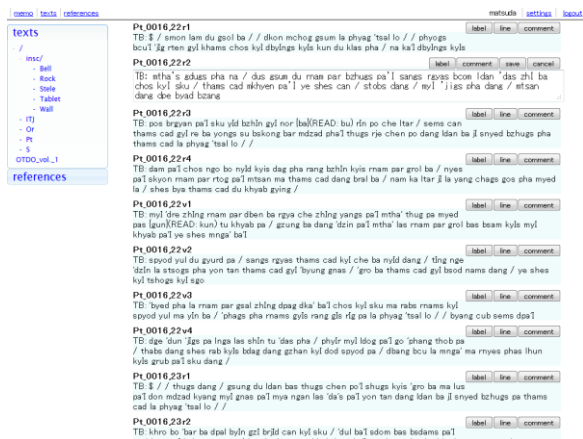


図 13: テキスト編集

識しているという点がある。

ただし，テキスト編集，あるいはそのメタデータの保存に関しては検索やチベット語処理にも幾分絡む問題であるため，次節で触れることとしたい。

7. 新方式の取り組み（2）

本節では，おおよそ方針の固まっている前節と異なり，試行段階の取り組みについて紹介したい。

■チベット語転写とチベット文字

先に挙げた

b4) チベット語転写の n/ng, s/sh, z/zh の区別は，チベット文字とその転写方式によるものである。

チベット文字あるいはより広くインド系文字には，子音に無気・有気音の別（k/kh 等），複数の鼻音，複数の歯擦音がある場合が多い。転写方式にもよるが，単一の子（母）音字が，複数のアルファベットを用いて転写されることもしばしばである。チベット語に関して言えば，おそらく最も広く用いられており，本プロジェクトでも採用されている（拡張）ワイリー方式によれば，鼻音として ‘nga’ ‘nya’ ‘na’ ‘ma’ があり，歯擦音には ‘zha’ ‘za’ ‘sha’ ‘sa’ がある。¹ これらをそのまま検索にかけると，n/ng, s/sh, z/zh の区別がつかないということになってしまう。²

これに対処する方法として，そのまま検索して適宜除外する方法，独自エンコードを用いる方法，Unicode のチベット文字に変換してしまう方法など，いくつか手段が考えられよう。

¹ インド系文字の場合，原則として母音/a/を含む音節文字であり，必要に応じて他の母音記号等を付す。

² チベット語の性質上 ‘ny’ が混同される状況になることは基本的にはない。

この中で Unicode 利用に関して少し言及しておく、チベット文字は一般に一文字から七文字程度の文字が組み合わさって一つのまとまりを作る。その中の一つが基字となって前後上下に文字がくっつく形であり、さらに母音記号がつくこともある。その形状に応じて転写が異なってくるということで、一般的なインド系文字よりは若干複雑な様相をもつ。したがって、単純な相互変換であっても少し煩雑な変換を要するが、近年ではサーバ側でもクライアント側でも処理性能は上がっているため、いずれにおいても敷居は下がっているといえよう。

また、近年チベット文字入力手段も一般的に利用されるようになり、Unicode で直接チベット文字を入力する傾向は増しつつあるが、一方で学会誌等ではまだアルファベット転写が大勢であるというのが現状である。

また、本プロジェクトが扱うテキストは碑文や写本といったものを含むため、必ずしも現代の標準的な表記におさまらないところもあり、さらに、部分的な、たとえば一つの子音要素だけの、欠落等があった場合でも、注記しやすいという意味でも、アルファベット転写を原則とし、多少の不可情報も落としても、必要に応じてチベット文字表示を付加する、という方針が妥当ではないかと考えているが、これに関してはさらに検討したいと考えている。

検索利用者にとっても単純な検索に限れば、アルファベット転写・チベット文字の両者をサポートする方が利便性は増すであろう。

■国際敦煌プロジェクト (IDP) との連携

OTDO では数年前から、IDP との連携を図っており¹、相互に緊密なリンクを行うことを同意している。

しかしながら、現状としては、OTDO 側では各資料の転写テキスト、IDP 側ではその元となる資料のカタログ情報と画像との間で、相互リンクがなされているのみである。²

これだけでも非常に有用な情報であることに間違いはないが、利用者視点で見ると、やはりものたりない感は否めないように思われる。

現在、OTDO サイドからのより緊密なリンクへの第一段階として、サムネイル表示を取り入れることを計画しているが、IDP 側のサイト構成の問題も若干あり、現状のまま取り入れることは若干躊躇される状況にある。

¹ たとえば以下を参照されたい。

<http://idpuk.blogspot.jp/2011/02/meeting-of-otdo-editorial-board-in-kobe.html>

² OTDO 側に関しては図 2 およびそれに関連する記述参照。IDP 側では “Click here for transcription of xxx at the OTDO website” 等としてリンクされている。

また一方で、OTDO 側としても、行単位のデータを保持しており、その記述に関するメタデータを作成しているのだから、そのようなデータを適切な形（たとえば XML あるいは JSON 形式）で提供する API 等の準備をしていくことも今後チベット学ないしその他関連するさまざまなプロジェクトとの連携をはかる上で重要になっていると考えている。

8. 実際の作業からのフィードバック

本プロジェクトは現在、基本的な実装のうち大部分を完成し、実際の作業を担当する研究者からのフィードバックを受けつつ、インターフェースを改良し、あるいは不足していた機能を追加するという作業を行っている。

その作業は残念ながら半ばであり、まだ十分な数もそろっていないので、そのすべての成果をここで報告することはできないが、簡単に触れておきたい。

全体的なコメントからは、本システムについておおむね好印象を与えていることが感じられる。また、問題点としても、現時点では、システムに直接関わる問題というよりは、データ構造の変更に伴う問題が多いように思われた。たとえば、先に紹介した階層ラベルの扱いで、実際の階層構造に合致していないので、修正する、とのコメントがあった。これは従来階層構造に関して明示的な形で表現できていなかった点が、階層ラベルの導入により、かえって問題が顕在化した例であるといえる。

9. 結び

以上、これまでの OTDO の取り組みと既存のシステム、それに対して新たなインターフェースのあり方を報告した。特に前稿では十分に紹介することのできなかった、チベット学における本プロジェクトの位置づけ、IDP あるいはその他プロジェクトとの連携を視野に入れた可能性、また、実際の編集者からのフィードバックに関しても少しく触れた。

本プロジェクトが 1980 年代という比較的早い段階からコンピューターを利用した研究を行い、構成人員としては小規模とはいえ、オンラインでの学術的共同作業を通してその成果を公開してきたこと、そしてその経験を踏まえた上で新たなインターフェースを構築しつつあることは一定の意義を有するものであると考えている。

今後も内実としてのテキストはもちろん、インターフェースおよび内部データ構造の両面での充実をはかることによって、国際的な取り組みにおける連携という面でも新たな役割を果たせるようになることが望ましいと考えている。

参考文献

- [1] 松田訓典, 星泉 : OTDO における古チベット語文書のオンライン共同編集作業, 情報処理学会研究報告, Vol. 2012-CH-94, No. 9, pp. 1-6, 2012.
- [2] Imaeda, Y., Takeuchi, T.: *Choix de documents tibétains conservés à la Bibliothèque nationale*, Tome III, Corpus syllabique, Bibliothèque nationale, Paris, 1990.
- [3] Imaeda, Y., Takeuchi, T., Hoshi, I., Ohara, Y. and Ishikawa, I.: *Choix de documents tibétains conservés à la Bibliothèque nationale*, Tome IV, Corpus syllabique, ILCAA, Université des Langues Étrangères de Tokyo, Tokyo, 2001.
- [4] Imaeda, Y., Takeuchi, T., Hoshi, I., Ohara, Y., Ishikawa, I., Iwao, K., Nishida, A. and Dotson, B.: *Tibetan Documents from Dunhuang, kept at the Bibliothèque Nationale de France and the British Library*, Old Tibetan Documents Online Monograph Series Vol. I, ILCAA, Tokyo University of Foreign Studies, Tokyo, 2007.
- [5] Iwao, K., Hill, N. and Takeuchi, T.: *Old Tibetan Inscriptions*, Old Tibetan Documents Online Monograph Series Vol. II, ILCAA, Tokyo University of Foreign Studies, Tokyo, 2009.
- [6] Imaeda, Y., Kapstein, M. T. and Takeuchi, T.: *New Studies of the Old Tibetan Documents: Philology, History and Religion*, Old Tibetan Documents Online Monograph Series Vol. III, ILCAA, Tokyo University of Foreign Studies, Tokyo, 2011.
- [7] 永崎 研宣, 苔米地 等流, Wangchuk, D., Almogi O., 下田 正弘: 人文学のためのコラボレーション—ITLR コラボレーションシステムの開発を中心的事例として—, じんもんこん 2011 論文集, 155-160. 2011.