chapter.12
サンスクリット文献電子データについての雑想

苫米地等流

1. はじめに

はじめに筆者のデジタルリソースに対する基本的なスタンスを明らかにしておくと、まったくもって身も蓋もない言い方ではあるが、テキスト校訂などの作業補助ツールとして便利であれば、手に入るものは精々ありがたく使わせてもらうまでのことであり、それ以上でもそれ以下でもない。よって本稿においても、これといって人文情報学 /Digital Humanities 的に目新しいことを述べるつもりはさらさらないし、その用意もない。また、サンスクリット文献のデジタルリソースについて筆者の観測範囲内にある程度の事柄は、昨今のサンスクリット文献学者であれば誰でも知っていると思われるので、いまさらここに書き記すことにさしたる意味があるとも考えてはいない。

とはいえ、エンドユーザー目線でデジタルリソースの現状についての雑感を書きとどめておくことにまったく意味がないわけでもあるまい。そこで、あくまで筆者自身の心覚え以上に出るものではない（*kevalam* ātmasmṛtaye）ことを一応お断りした上で、サンスクリット文献資料を取り巻くデジタルデータの状況についてとりとめのない雑文をものして形ばかり紙幅を費やすこととする。なお以下では、（いちいち面倒なので）リソースの URL などの情報は一切割愛するがよろしく了とされたい。そもそも、ウェブ検索で大抵の参照先を発見可能なのがデジタルのメリットでもあろうから。

2. 電子テキストデータベースかリポジトリか

漢訳大蔵経やチベット大蔵経・蔵外文献のようなある程度まとまった分量と体系を持つコーパスの場合とは異なり、現在利用可能なインド文献の電子テキストは、SAT や ACIP、BDRC に代表されるような大規模プロジェクトによるものではなく、個々の研究者が自己の関心に応じて入力したものや比較的小規模な研究プロジェクトの成果物として作成したものが中心となっている。従って、それら電子テキストには、フォーマットや品質においてばらつきが生じるのは避けがたい。また、このようなばらつきを解消して、統一的な様式を持つまとまったデータベースに構成しようという動きもあまり見られない。そもそも、現在みられるインド文献の各種デジタルリソースは、検索インターフェースを備えた高度なデータベースというよりも、雑多な電子テキストをまとめて保管しておくリポジトリとしての性格が強いものが多い。この種のいわばプリミティヴなリポジトリは DH 的には面白味に欠けるかもしれないが、別に DH 的関心を優先させねばならないという法もない。ユーザーの視点からすれば、ややもすればリソース提供者の自己満足の域を出ない「気の利いた」機能を持つデータベースよりも、ローカルのストレージにダウンロードした上で自身のやりたい方法で好きなように利用できるデータの方がはるかに有用である場合も少なくないだろう。実際、筆者の周囲のサンスクリット研究者の多くは、リポジトリあるいは個人的なルートを通して入手した複数のデータファイルを、パソコンのターミナルエミュレータ上で grep などのコマンドラインツールを使って一斉サーチするというかたちで利用している。文献研究を行う上では、このような利用法でも取り立てて不都合はないし、データさえ手元にあれば別にウェブ上のデータベースで検索できなくてもまったく困ることはない。データ品質のばらつきについてもあらかじめ織り込み済みで、わかった上で利用すればよいことである。

3. 還元梵文によるデータのコンタミネーション

研究リソースとしての電子テキストに重要なのは、データの質やインター

フェースよりも、まずは物量である。多種・多分野のテキストがデータとして大量利用できるようになることの恩恵は大きい。しかし、その一方でまったくのゴミとしか言いようのないデータがかなりの量流通しているのも事実である。これはデータの質うんぬん以前の問題で、そもそもサンスクリット原典の現存が確認されていない文献が「サンスクリット」電子テキストとしてリポジトリに保管され、データのコンタミネーションを起こしている現状があるのである。現在最も広く利用されているであろうインド文献電子リポジトリであるゲッティンゲン大学の Göttingen Register of Electronic Texts in Indian Languages（GRETIL）は、ヴェーダから仏教文献にいたるまで広範なジャンルのデータを網羅的に収集しているが、それらのうち仏教文献のかなりのものがいわゆる「還元梵文」のデータなのである。インド仏教研究のコンテクストで蔵漢訳仏典を読む際に背後にあるサンスクリットを想定する必要があるのは当然である一方、文献全体を蔵漢訳から還梵するという行為は、註釈など関連文献から十分な量のサンスクリット断片を得られる場合を除いて学術的にはほぼ無価値である。欧米では還元梵文の出版はほぼ行われなくなったが、残念ながらインドなどでは――サンスクリットに対する国粋主義的プライドの所為でもあろうが――現在でも還元梵文が学術成果として出版され続けており、これらが電子データ化されてウェブ上に流通、GRETIL のようなリポジトリに収録（混入？）される事態が発生しているわけである。GRETIL に収録されている仏教関連の電子データの中には、University of the West の Digital Sanskrit Buddhist Canon（DSBC）プロジェクト由来のものが多く含まれるが、この DSBC こそがデータ汚染の最大の発生源と言ってよいだろう。例えば、GRETIL には、DSBC 由来のデータとしてアティシャ／ディーパンカラシュリージュニャーナ著作の「サンスクリットテキスト」が 11 点収録されているが、アティシャ著作のサンスクリット原典はまだ一点も確認されていないことはインド仏教研究者であれば誰もが知るところである。ほかにも、ディグナーガの『観所縁縁論（かんしょえんねんろん）』の「サンスクリットテキスト」などというびっくりするようなデータも含まれていたりするが、これらの大半が DSBC 由来のものなのである。もちろん、DSBC も素性の正しいサンスクリットテキストの電子化に大きく貢献してはいるが、プロジェクトの性格が純粋に学術的というよりは、宗教的ミッションを動機とし

ている部分が大きいためか、電子化すべき資料の取捨が甘くなっていると考えられる。一度ウェブ上に掲載されたデータは、検索エンジンによってインデックス化され、その後の検索結果に影響を及ぼす。結果、オーセンティックなテキストも、還元梵文も区別されることなく検索結果として表示され混乱をもたらすことになる。また、grep などでサーチを行う場合も、検索結果の信頼度を下げることとなる。現在 GRETIL では、収録されている電子テキストを zip ファイルとして一括ダウンロードできるようになっている（以前は wget などのツールを使って再帰ダウンロードする必要があった）が、これを学術研究に利用する際には、まず還元梵文データの草むしりをするなど、ひと手間かける必要がある。ただし、GRETIL のファイル命名規則（MS-DOS 時代の 8+3 文字制限が残っている）からはファイルの内容を直感的に知ることは困難であり、いちいちファイルを開いて判断しなければならない。今後は、GRETIL から素性の怪しいデータは削除あるいは分離し、信頼できるデータのみを一括でダウンロードできるよう改善してもらえればありがたい。

4. 最後に、テキストの構造化について

データベースでなくても簡単なリポジトリで実用上はとりあえず十分だとは思うものの、今後作成される電子データはやはり体系的なデータベースへの収録を見据えた構造化データであることが望ましいには違いない。GRETIL もデータの TEI 準拠した構造化を順次進めていく方向のようである。より本格的なインド文献のデータ構造化プロジェクトとしては、SARIT がよく知られており、SARIT の関係者が中心となって TEI ガイドラインへインド古典文献学のニーズを反映させるための分科会 SIG Indic も立ち上げられている。とはいえ、サンスクリットは、連声や複合語形成、音節構造と語の関係などの言語的特徴から、ネスト構造を基本とする XML によるマークアップに対する障壁が大きく、SARIT や SIG Indic の試みがどの程度状況を改善するかはいまのところ未知数というべきであろう。筆者も、テキスト校訂の基礎データには TEI 準拠の XML を採用してはいるが、上記の問題には基本的に頬かむりして、とりあえず紙媒体（PDF 含む）への出力へ変換できればよしと割り切って使っ

ている。また、タグによって構造化されたデータは、従来電子テキストが使われてきたやり方、すなわち grep などのツールによるサーチとは相性の悪い部分がある。タグでテキストが分断されることで、検索の精度が大きく下がるわけである。これを解消するには、eXist や BaseX などの XML データベースを運用するといった方法があるが、多くの研究者にとってこれは、いまのところ敷居の高い選択肢と言わざるを得まい。結局のところ、一番使い勝手のよいデータはプレーンテキストという身も蓋もない話になってしまうが、筆者はこれはあながち無視できないことだと考えている。研究者がプレーンテキストを必要としているのなら、そのニーズを大切にするのがリソース提供者として必要なスタンスであろう。もちろん、DH 的に高度な設計がなされたデータベースを提供することも必要であるが、それが利用者に対して使い方や方法論を押しつけるようなことになっては本末転倒である。研究基盤のデジタル化が仏教学・インド古典学を含む人文学を真の意味でドラスティックに変容させるか否かは筆者にはわからない（と、一応韜晦しておく）が、いずれにせよ、DH に携わる人々にはあくまで謙虚に、利用者ファーストでリソース開発してほしいものである。