

(2) 新しいアイヌ語辞書へ

また、新しいタイプのアイヌ語辞書の登場が期待される。紹介した通り、単語を検索すると音声つきの例文がどんどん蓄積される。また現在もアイヌ語テキストは単語レベルでデータベース化し、言語学者が文法情報（日本語クロス）をつけ始めている。また、既刊のアイヌ語辞書を電子辞書化し、同じサイトで公開を開始した。今後言語学の優秀な研究者が加わることでこれらが統合され、今までにないアイヌ語辞書が生まれることが期待される。

(やすだ・ますほ／アイヌ民族博物館学芸員)

第41回大会シンポジウム「口承文芸デジタルアーカイブの課題と展望」

音声等を含む多様なデータベースの現状と課題 —人文情報学の立場から—

後 藤 真

本報告では、特に人文情報学の立場から、口承文芸に関わるデータベースの一般的課題について述べるとともに、国立歴史民俗博物館（以下歴博）が保有する民謡データベースの概要について説明するものである。

特に、本報告では、民謡や音声、口承文芸に関わるデータをいかに長期的にアクセス可能な状態とし、かつ研究するものとして使うことができるかについて述べる。この課題を整理するためのひとつの材料として、また情報の提供として歴博の「民謡データベース」を紹介したい。その後、人文情報学の現状と、それを踏まえたさまざまな口承文芸のデータベースの課題へと戻ることとする。

「民謡データベース」は、国立歴史民俗博物館で、館内限定で公開されているデータベースである。データ件数は69000件、文化庁の民謡緊急調査（1979～1989）をもとにしたものである（図1）。基本的には関連するメタデータと、音声データを収録しており、2007年より館内公開を開始した。メタデータ項目は表1のとおりである。

当館内田順子によると、本DBは著作権等については、基本的に研究利用に限り許諾を得ており、データの公開については問題が少ない。原則として、研究者限定の公開であり、その根拠は各都道府県及び文化庁との間で取り決められた「国立歴史民俗博物館 全国『民謡緊急調査』資料活用事業実施要領」によっている。館内のデータベースで必要な情報を検索し、その上で特定の音源等にアクセスしたい場合には、CDにて聞くことが可能である。このCDは各都道府県が作成した録音テープから作成したものであり、これに

外に長期保存等を目的とした DAT の複製もある。なお、館内公開に限っているのは、民謡の音声内容にセンシティブな内容が含まれている可能性が高いが、その内容がどこにあるか必ずしも判然とせず、結果的に全面的な公開が困難となってしまっているためである。

しかし、館内公開では、当然利用者が少なく、当時の貴重な音源を活かしきれていない状況になってしまっているのも事実である。センシティブなデータの中から、公開可能なデータを効果的に導き出し、これらの資料をさまざまな研究などで用いてもらうようにするか、まさに検討が始まったばかりである。また、文化庁における「民謡緊急調査」時の録音テープが劣化してしまい、地方公共団体から音源の複製依頼なども来ており、これらに対応することも、民謡データベースの役割として求められている。このようにもともとが機械可読メディアである場合には、おおとのマスター自体の劣化が激しくなり、さまざまな処理が困難になることがある。そのため、デジタル複製を別の場所に置いておくこと自体が、データベースの重要な役割になることもある。

それでは、あらためて人文情報学の概要に戻ることとする。そもそも、人文情報学とは、人文学のあるテーマの解明手法として情報技術を用いる学問である。人文情報学という学問そのものは日本においては情報学の分野からスタートしたものが多く一方で、欧米、とりわけ英米圏では人文学、中でも文学研究の中から生まれたものが多かった（英語では Digital Humanities と訳される）。特に Digital Humanities という言葉としては、日本では 2000 年代後半ぐらいから積極的に取り入れられている。また、あわせて日本では文化に関する資料をデジタル化する際には「デジタルアーカイブ」という言葉も多く用いられてきた。当初（2000 年代前半）はこの「デジタルアーカイブ」は、長期的な資料の維持・保存という本来的な「アーカイブ」や研究の機能が重視されることなく、鑑賞を中心とするものであった。その後、人文情報学の発展にともない、これらのデジタルアーカイブが研究基盤として用いやすいかたちで構築されるようになってきた。それは、Web の技術の進歩によって長期保存・共有を可能とできるデータベースの構築が可能となった面が大きい。そこで、これらの長期保存を可能とするデータについて整理し、口承文芸関係のデータベースの課題を導き出す。

1. 文字情報に関するデータベースの保存

文字情報に関するデータの長期保存については、基本的には安定しつつある状況である。大本となるテキストはフォーマットが安定したといえる。また、文字コードについても UTF-8 といわれる形式が確定的になってきており、文字コードの問題でのちに読めなくなるという事態は避けられそうな情勢である。また、版面データは（使い勝手に問題はあっても）PDF がいわゆるデファクトスタンダードの位置を占めているといえるであろう。

日本民謡データベースの検索

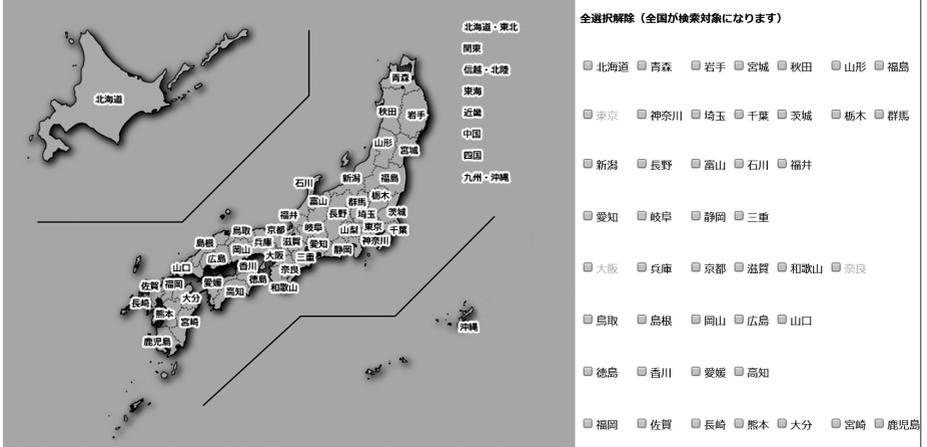
検索語を入力し、【検索】ボタンを押してください（[検索語例](#)）
 本データベースのご利用は研究目的に限らせていただきます

結果表示件数:

フリーワード:

大分類:
 中分類:
 小分類:

都道府県（複数選択可・何も選択しなければ全国が検索対象）:



北海道:東北
 関東
 信越:北陸
 東海
 近畿
 中国
 四国
 九州:沖縄

全選択解除（全国が検索対象になります）

北海道 青森 岩手 宮城 秋田 山形 福島
 東京 神奈川 埼玉 千葉 茨城 栃木 群馬
 新潟 長野 富山 石川 福井
 愛知 岐阜 静岡 三重
 大阪 兵庫 京都 滋賀 和歌山 奈良
 鳥取 島根 岡山 広島 山口
 徳島 香川 愛媛 高知
 福岡 佐賀 長崎 熊本 大分 宮崎 鹿児島

項目一覧表示:

個別項目検索:

名称

地名

分類

図1 日本民謡のデータベースの検索画面

2. 画像情報に関するデータベース

画像情報に関しては、長期保存に関しては tiff による保存がスタンダードになりつつあるといえる。また、ここ1~2年ほどで IIF (International Image Interoperability Framework) という画像の相互運用の仕組みも完成しつつあり、比較的安定したものになっている。

これらのデジタルデータの長期的な保存と公開の課題は、もっぱらそれらのデータを格納する「メディア」にあるといつてよい。本稿執筆時点で最も読み出しやすい HDD やフラッシュディスク (USB メモリや SSD など) は、その寿命が紙などに対して長いとは言えない。また DVD などの光ディスクは媒体としては長持ちするが、これを読む機械その

ものの寿命が長いとはいえないので、それらの点を考慮し、データの移行を随時行っていく必要がある。

3. 音声・動画に関するデータベースの保存

一方で、課題が多いのは音声や動画コンテンツに関するデータである。フォーマットについては、デファクトスタンダードとして、音声の公開部分についてはMP3がこれまで優勢であったが、今後はAACなどへと転換するという可能性も指摘されており、必ずしも安定的なものとは言えない。保存についてはWAV形式が安定的なポジションとなりつつあるといえるであろうが、安定的なものになるまでにいましばらくの時間を要するといえるであろう。動画についても、フォーマット形式が今後の課題となる。

つまり、口承文芸に関するデータは、目録や文字おこしたテキストに関しては長期的な提供が可能である。一方で、これらの目録と関連するような形で動画や音声を入れる場合には、なお注意が必要であるといえる。

次に、これらのデータの公開について述べる。テキストデータに比べると、音声や動画の権利処理は複雑である。動画や音声の中から、不適切な部分を取り除くためには、実際に「それが不適切である」とわかる

人間が、確認しなければならない。その際、30時間のデータがあると30時間分確認する必要がある。動画の目視であれば、早送り等が可能であろうが、音声データを聞き取る場合にはその早送り等にも限度がある。また、動画の場合には著作権だけではなく、肖像権などの

【分類】		
【大分類】	【中分類】	【小分類】
	踊り歌・舞謡	盆踊りの歌
【市町村コード】		12212
【市町村名】		千葉県 佐倉市 (さくらし)
【集落名】		江原
【名称フリガナ1】		ムギツキウタ
【名称1】		麦搗き歌
【名称フリガナ2】		—
【名称2】		—
【名称フリガナ3】		—
【名称3】		—
【歌う機会】		盆
【歌う日時】		—
【歌う場所】		—
【歌う人の性別】		—
【年齢層】		—
【資格の有無】		—
【資格の内訳】		—
【録音の歌の編成】		1人
【報告書の歌の編成】		—
【録音の楽器の有無】		無
【報告書の楽器の有無】		—
【録音の楽器及び数】		—
【報告書の楽器及び数】		—
【録音データの有無】		有
【収録年】		S54
【演唱時間】		1:09
【その他の音楽的情報】		—
【報告書の音楽的情報】		—
【歌に伴う身体の動き】		—
【詞型】		—
【詞型の内訳】		—
【歌詞】		—
【演唱者1】		明治41男
【演唱者2】		—
【演唱者3】		—
【演唱者4】		—
【演唱者5】		—
【報告書及び調査票による情報】	H	
【録音時等の情報】		—
【DAT】		12-012/5319
【報告書頁】		—

表1 民謡DBの項目

問題があるため、より複雑な構造がある。

また、音声や動画データは検索が難しいのも課題である。テキストデータは、データの公開が難しい部分を検索し、確認して削除するなどの作業が可能となる一方で、動画や音声はこれらの検索が困難である。したがって、公開のためのチェックのハードルはより高いものとなっている。当館の民謡データベースも同様の課題を抱えているといえるであろう。

一方、これらを解決できると大変に魅力的なものになるのもたしかであり、これらを公開することによるインパクトは非常に大きいことも強く主張しておきたい。

デジタルデータの公開は、常に資料の可能性とチャンネルを広げ、多くの人に資料とその研究の価値を届けるものである。このように述べた際、デジタルデータを公開することでアナログデータが駆逐されてしまうという懸念、もしくはデジタルデータによることでアナログデータは不要であるという大きな期待の2つが出される。しかし、基本的にはデジタルとアナログは対立的なものではないことを主張しておきたい。両者は補完的な存在である。紙は、現時点においては、デジタルデータやメディアよりは相対的に長く持つが検索性という観点では決して優位とは言えない。一方、デジタルデータは検索性やアクセスするチャンネル、共有が容易だが全体を見るという点や長期保存ではまだ課題が残っているといえる（※ただし、テープ類のようなアナログでも機械でよむことが前提のものはデジタルデータと同様のもろさを持っているので、十分に注意する必要がある）。したがって、デジタルメディアに変える＝アナログを駆逐するというのではない。

デジタルデータを蓄積し、共有することは、アナログにとってもきわめて重要なことである。特にデジタルで識別できるようにすることによって、本体を長く残す工夫の一つともなるのである。一般的に広く認知されている資料は、多くの人目があるので、誤って廃棄されたりすることも少なくなり、残りやすい。発見可能性をより広げるためにもデジタルデータの蓄積と公開は重要なことであると言えよう。

国立歴史民俗博物館でも、昨年度より「総合資料学」の創成という事業の中で、歴博や大学・博物館等にあるさまざまな資料を、新たな形でデータ化する試みを開始している。これにより、より汎用性の高いデータ形式で、広く発見されるデータ群を作り、大学歴史研究資料・博物館資料の全体のネットワーク構築することを目指している。このような事業の中においても、民謡データベースについて取組を工夫していきたいと考えている。

デジタルデータは、いきなりすべてを解決する夢の道具ではない。一方で、私たちの資料を破壊するための悪魔の道具でもない。デジタルデータの優位性を十分に活かすことで、これまででは困難であったデータベースも構築可能となり、研究や教育、口承文芸に触れる文化などこれまでに「加える」かたちでさまざまな広がりを持つものであろうと考えている。

(ごとう・まこと／国立歴史民俗博物館)