ことばの研究資料アーカイブ
—国立国語研究所所蔵資料の利活用に向けて—

寺島　宏貴　　星野　雅英　　高田　智和
国立国語研究所　　同　　同

東京都立川市に所在する国立国語研究所（NINJAL）は，日本語研究に関する様々な資料を作成・蓄積してきた．本稿では，研究所に所蔵される紙・メディア資料を紹介するとともに，これらの資料を管理する際の課題やXMLの記述形式による旧データベースの問題点を指摘する．また，現在行なっている新データベースに向けた基礎作業，および今後取り組むべき課題について述べる．

Practical using material owned by National Institute for Japanese Language and Linguistics (NINJAL)

Hirotaka Terashima　Masahide Hoshino　Tomokazu Takada
NINJAL　NINJAL　NINJAL

National Institute for Japanese Language and Linguistics (NINJAL, Tachikawa-shi, Tokyo Metropolis) have been accumulating research documents and media for linguistics (e.g., phonetic data recording on cassette and digital audio tapes) .This paper attempts to explain contents of these materials owned by NINJAL. A further important point is creating database for these abundant materials. We recognize the advantage of relational database for description of research materials and contents about past research project is better than XML which has used in our laboratory.

１．はじめに

国立国語研究所（NINJAL=National Institute for Japanese Language and Linguistics, 東京都立川市）は1948（昭和23）年12月に創設され，独立行政法人を経て，2009（平成21）年10月1日に大学共同利用機関法人 人間文化研究機構の 6 番目の機関として設置された．

新たに設置された国立国語研究所は，これまでの研究の蓄積を踏まえながら，研究の国際的拠点として国内および海外の大学・研究機関と大規模な理論的・実証的共同研究を展開することによって日本語の特質の全貌を解明し，言語の研究を通して人間に関する理解と洞察を深めることを目的としている．また，共同研究の成果や関連する研究文献情報を広く社会に発信・提供し，日本語教育，自然言語処理など様々な応用面に寄与することも重要な使命としている．

これまで国語研は 70 年近くにわたって方言調査や語彙，音声，日本語教育，言語コーパスなどといった多様な研究プロジェクトを積み重ねてきた．過去に実施された調査研究では，音声・映像・電子メディアを含む膨大な研究資料が生み出された．しかし，プロジェクトの成果である学術論文や刊行物は別として，国語研の所蔵資料が研究所の内外で利活用される機会は少ない．

本発表では言語学ないしは日本語学という，ことばを扱う領域にとっての研究資料アーカイブはいかなる形が望ましいかとの課題を念頭に，国語研で現在実施している，アーカイブ構築に向けた取り組みを報告したい．

具体的には，研究所の所蔵資料について簡単に紹介したのち，これまで国語研において行われてきた研究資料のデータベース化について述べるとともに，その問題点を指摘する．ここでは資料移管時の過誤や，データベース化に際しての資料群記述における問題を扱う．特に，これまで国語研で採用していたXMLデータベースの問題点を述べたい．

最後に，以上を踏まえた今後の検討課題について述べる．まず研究資料のデータ管理としては表形式のリレーショナルデータベースが望ましいこと，ならびに新データベースに向けた基礎作業について説明する．さらに，電子媒体を含む多種のメディアを今後どのように管理していくべきか，また閲覧環境の整備によって外部（他の研究領域）への利用提供をどう図るのかを議論したい．

２．国立国語研究所所蔵の各種資料

２－１　紙資料

国語研の所蔵資料は，主として紙資料と各種のメディア資料とに大別される．過去の研究プロジェク

トにおいて使用されてきた機器類も保管しているが，ここでは紙とメディアのみについて述べる．

まず紙資料の一例として，社会調査において作成された調査票の原本やその集計・分析資料や，語彙調査の対象となった学校教科書や雑誌などの資料現物が保管されている．

また，各研究プロジェクトの遂行にあたって作成された公文書等の事務書類も散見される．これらは文書箱に収納し，研究プロジェクトごとにグルーピングした上で，紙資料専用の収蔵庫（中央資料庫）に保管している．中央資料庫は，資料保存の観点から温度を23℃に一定して管理されている．

現時点での文書箱の総数は約4,000箱程度であるが，今後も所内各部署からの資料受入によって収蔵量の増加が見込まれる．これに加え，方言や語彙の調査で作成された推計約1,000万枚のカード類が木製・ステンレス製のカードケースに，同じく地図資料も専用ケースに保管されている．

以上に述べた紙資料の中には,昭和20年代の社会調査・語彙調査に関する貴重書類も見られる．

２－２　各種メディア資料

次にメディア資料は，各種媒体に記録された調査・分析データが大量に所蔵されている．中でも，過去の言語調査において録音・採集された音声が膨大にあり，特にカセットテープ（ct），DAT テープ（dt）に記録されたものが圧倒的多数にのぼる．

約50,000本のct，dtには，社会調査での外国人を含むインフォーマントによる会話・談話を始めとして，語彙調査等のデータ採集用に作成されたテレビ・ラジオ・各種講演の録音がある．なお国語研が各地で開催した研究会や,国際シンポジウムの録音記録も残っている．

この他にも，CD，DVD，MD，オープンリールテープ，オープンリールビデオ，8mmフィルム，ß・VHSビデオカセットテープがある．またメディアには音声・映像ばかりではなく，調査研究の過程で作成された磁気テープ，フロッピーディスク，MOディスクといった媒体も含まれる．

以上に紹介した各種メディアはメディア保管庫に収納し，紙資料と同様23℃で管理されている．

３．データベースについて

３－１　これまでのデータ形式

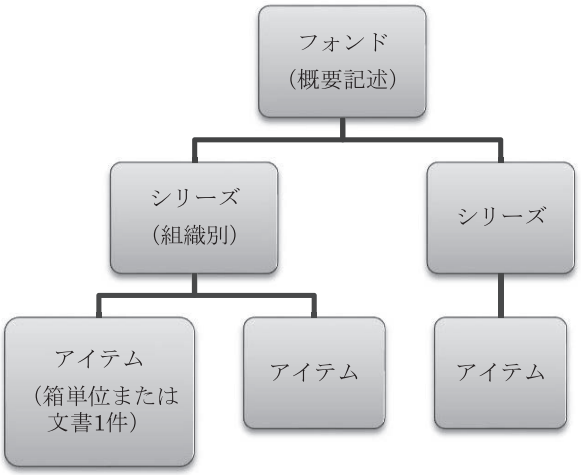
次に，これまで国語研で行なってきた研究資料のデータベース化と，その問題点を説明する．国語研では下記の流れで，資料のデータベース化を図ってきた．

① 移管者の側で資料の移管票を作成し，資料受取者へ提出．
② 資料の受入および燻蒸．
③ 燻蒸後の紙資料・各種メディア資料をそれぞれの保管庫に移管．
④ 移管票の情報や資料現物をもとに，EAD（Encoded Archival Description アーカイブ符号化記述）によって，各資料群の概要についてXML記述を作成する．

このうち④のEADはXML一般ではなく，電子的検索手段の特性に配慮された文書記述言語 SGML（Standard Generalized Markup Language）サブセットの国際規格である．

EADのメリットとしては,アーカイブの群や小群からなる階層的な秩序を反映できることに特性がある．これまで国語研での EAD 運用に際しては，資料の移管後，配架された資料（紙資料・メディア）に基づいて概要把握を行なった．

▼EADのツリー構造（国語研の場合）



下記のように，資料を１つのフォンドと複数のシリーズに階層分けし，必要事項を記述する．原則一つの研究プロジェクトを１資料群として，プロジェクトの課題・研究内容や組織歴に基づき，次の各階層・記述項目によってXML記述を作成した．

(1)フォンド（fond）レベル

共通の出所を持つ資料群の全体（レコードグループ）．原則，一つの調査・研究プロジェクトに相当する．下記項目の記述を行なった．

・版（第n版）
・記述名称
・発行者
・記述作成者
・発行年月日
・大きさ
・形態／媒体
・表題（調査名，研究プロジェクト名）
・要約（資料群の要約）
・資料作成年月日（現状では年代のみ）
・配架位置（中央資料庫，メディア庫など）

・管理歴（移管元と移管先）
・資料作成者（担当部署と担当者）
・組織歴，履歴
・入手情報
・数量，形態（紙資料とメディアの数）
・物理的特徴
・編成（階層分け）
・資料内容
・評価
・追加受入
・関連資料（別フォンド等）
・分離資料
・公開条件
・利用／複写条件
・必要機器（記録媒体に関する事項）
・他の検索手段
・オリジナル資料
・複製
・関連出版物（成果刊行物およびメディア等）
・資料管理概要（現在の管理状況等）
・備考（資料の状態やメディア ID 等，留意事項）

(2)シリーズ（series）レベル

内部組織（ex.各研究ユニット）単位である．資料の作成部署や資料内容，もしくは小プロジェクトをもとにシリーズ分類を行う．記述項目はフォンドレベルより簡略なものとなっていた．

シリーズレベルの記述項目は次の通りである（一部）．

・表題
・資料作成年月日（年代のみ）
・配架位置
・資料作成者（特記すべき場合のみ）
・編成（サブシリーズレベル）
・備考（メディア ID 等）

(3)アイテム（item）レベル

資料 1 点ごとの最小分割単位である．項目はシリーズレベルより簡略なものとした．やはり上位レベルとの重複項目が多かったが，アイテムレベルのみ「収納容器」の項目が追加されている．

３－２　XML 階層記述のデメリット

次に，EAD 記述における問題点について述べる．

EAD の導入時には，おそらく多様な資料群に対応すべく記述項目が数多く設けられた．しかし，上記に挙げたように各階層は項目が多すぎ，上位・下位レベルでの重複もかなりあった．どの項目に，どのような情報を記述していくかが不明になりがちであり，同じ記述を何度も繰り返してしまうリスクが常にあった．また，アイテムは文書 1 件や各種メディア 1 本ずつ大量に登録されるケースがみられた．

これらの問題のため，記述者によって情報量にバラつきが出ることになった．さらに，記述者が下位のシリーズ・アイテムの階層を増やしたためツリーが細かく枝分かれした．場合によってはサブフォンド（subfond）レベル，サブシリーズ（subseries）レベルの記述も行うことがあった．後に，原則としてサブフォンド，サブシリーズ，アイテムの階層は使用しないこととなった．

このような各レベルの細分化のほか，大量のメディアが歴史的古文書の目録を作るように 1 点ずつ登録されることもあった．EAD の運用では，ベースとなっている移管票に情報を過剰に付加する形で，移管票よりもはるかに長大な記述が生まれた．そのため，各資料群の記述は，概要を示すにも関わらず概要把握には向かないという問題があった．

また２.で述べた中央資料庫・メディア庫に収納された資料については，移管票の他に，すでにそれぞれ表形式の目録が備わっている（詳細は４で述べる）．しかし，EAD 記述はこれらのリストとは全く別個に作成され，既存の目録とは無関係のデータベースとして運用された．

このため，既存の移管票と目録とがリンクしたデータベースを新たに構築する必要性が生じた．

４．今後の検討課題

４－１　既存目録を活用したデータベース（案）

次に，今後 EAD に代わる新たなデータベースの形，及びその課題点について述べたい．

国語研では現在，移管票と紙資料・メディア資料の目録（国語研では「配架リスト」・「各種メディア ID 一覧」と呼んでいる）とがリンクした，表形式のデータベース構築を検討している．

移管票に移管者・資料受取者が記入する項目は次のようになっている（5 ページの図も参照のこと）．

A　移管資料の概要
① 移管年月日
② 移管量（保存箱・メディア箱・カードケース箱ごとに数量を記載）
③ 移管者名
④ 受取者名
⑤ 調査・研究名称
⑥ 担当部署・担当者名
⑦ 研究・調査の概要（成果刊行物・関係調査等があれば記載）
⑧ 保存・利用にあたっての留意事項（資料の媒体・個人情報の扱い等）
⑨ 調査 ID（=fond ID．受取者の側で記載）

B　移管資料リスト（カード資料を除く）
① 保存箱（メディア箱）番号
② 内容（メディアの場合，その種別）
③ 所在（資料庫内の配架位置）
④ 備考（メディアの場合のみ）

このように移管票のフォーマットは，移管資料の概要とともに，文書箱 1 ケースずつの情報が記載可能なものとなっている．資料管理者は，上記 B を配

架リストに反映する．

ただし，現在の配架リストの項目は移管票に対応しておらず，下記のようになっている．

① 棚位置（配架位置）
② 作成部署
③ 調査名
④ 資料内容（文書箱 1 ケースの内容）
⑤ 箱 ID（fond no.＋box no.）
⑥ 移管者
⑦ 備考

他方，各種メディア ID 一覧に関してはアイテムレベルで ID を付与している（例えばカセットテープは ct0001）．この一覧ではメディアの種別ごとにシートが分かれており，入力項目もそれぞれ異なる．

以上のように配架リスト・メディア一覧・移管票は，ともに関連性の薄い形で整備されている．

今後，こうした現状を改めていく必要があるが，配架リストの記述レベルに関してはフォンド（レコードグループ）―シリーズ―アイテム（箱単位）と，すでに階層的な記述レベルに分類されている．この記述レベルに対応した記述要素のうち，上記配架リストの①～③がフォンドに相当する．新データベースでのフォンド階層は，検索時にキーワードの組み合わせや絞込ができない場合，目次的な「資料ガイド」の機能を担う．

上記④は，シリーズおよびアイテムに当たる．資料ガイドの下位レベルであり，研究組織に基づいた個別の分類を表す．箱単位のレコードのため，④の現状は「○○（シリーズに相当する）関係資料」のような記載があったり，かなり詳細に記載したりするなど精粗がみられる．これは資料現物の状態によるところが大きい．したがって，アイテムは文書 1 件ごとではなく，箱単位で上位レベルと結びつける必要があると思われる．

これら各階層のリレーションシップは，片方のテーブルの 1 情報（行）に対して，もう片方では複数の情報が関連付けられた「1 対多」である．検索のためには，1 対多の階層構造で結びついた資料目録を，階層的な画面体系で表現する必要がある．また，国語研ではこのようなデータベース構築の経験はないため，同じ記述を何度も繰り返さないことを目的とする正規化に際して，所内職員の共通認識を得やすい仮想表の設計から始めることが日々の業務と密接し，より理解しやすい設計手順となると予想される（大城 2000 参照）．

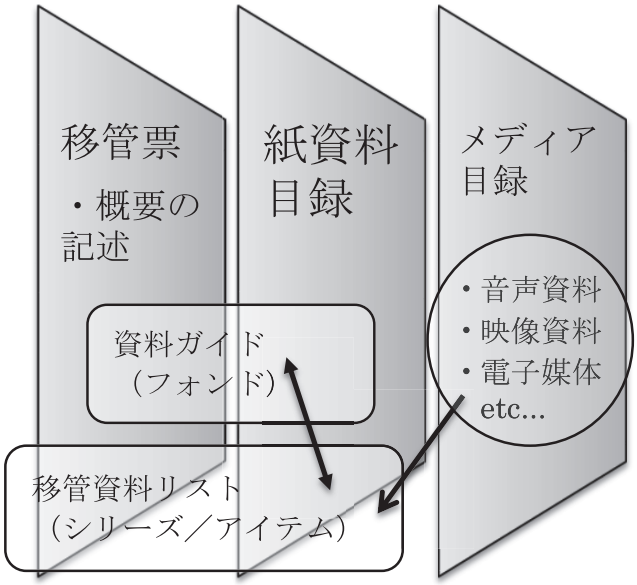
このように，既存の階層関係を生かしながら，手持ちの移管票・目録を活用したリレーショナルデータベースが研究資料の管理としては適切である．

４－２　移管記述の見直しと目録の限定公開

次に，新データベースに向けて所内で実施している作業について説明する．

これまで，EAD 記述のデータベースは国語研ウェ

▼移管票・資料目録を通じたリレーショナルデータベース（概念図）



▼紙資料目録（配架リスト）の階層関係．大城（2000）所載表を改変．

フォンド(レコードグループ)
シリーズ
アイテム

棚位置	作成部署	調査名	内容	箱ID	移管者	備考
002-C-1R	コーパス開発センター	日本語書き言葉コーパスの構築(BCCWJ)	台帳(書籍) BCCWJのサンプリング用台帳。	fo0208-02	○○	
002-C-2L	〃	〃	〃	fo0208-03	○○	
002-C-2R	〃	〃	〃	fo0208-04	○○	
002-C-3L	〃	〃	〃	fo0208-05	○○	
002-C-3R	〃	〃	〃	fo0208-06	○○	

ブサイトで公開されていた．しかし積極的な発信を行なっておらず，国語研の所蔵資料にどんなものがあるかは資料管理者のみぞ知る，という状況となっていたことは否めない．実際のところ，所内の職員や所外の OB・OG を中心に，ウェブサイトを通じた閲覧請求は年に数件であった．

それと同時に，利用者のための閲覧室も完備しておらず，所内図書館と比較しても研究資料の利用に適した環境の充実を図っていなかった．さらに，中央資料庫・メディア庫や，配架リスト・メディア一覧の存在すら所内で周知されているとは言い難い．

2013（平成 25）年 10 月以降，所内の組織替えに伴って新たな資料管理者に交替してから，利用者が研究資料にアクセスしやすい環境の整備を行なっている．具体的には，中央資料庫・メディア保管庫のインスペクション（棚卸し作業）を実施することによって，配架済みであっても配架リストに未入力であった資料の抽出や，移管票提出済み・燻蒸済みにも関わらず未移管の資料の特定を行なった．

燻蒸済み資料の未移管問題は資料管理者の側で発生した，アーカイブにおけるヒューマンエラーといえよう．ここ数年にわたって，専従の資料管理者を置かなかった点もその要因の一つであろう．

また，移管資料についてはこれまで全点を燻蒸処理していた．しかし，取り扱いに注意を要するような貴重書類は今後の移管資料には含まれない点を踏まえ，原則，燻蒸は実施しないこととした．

この一方，上記の資料庫インスペクションを経て，所内職員を対象に，配架リストを公開した（目録上の個人情報等は除外）．また，EAD 記述によって作成した資料概要については，記述項目の再選定を行い，移管票に準じた簡易記述に改めた．

5．おわりに

以上のように，ことばの研究資料のアーカイブ化にあたっては貴重資料を除くとしても歴史的文書の扱いとは異なった処置を採るべきである．これまで紹介した取り組みを現在行なっているとはいえ，新データベースの構築や，所蔵資料の管理方法についての具体的な見通しは未だ立っていない状況である．

資料情報の公開については概要記述から資料目録の重視へと方針転換した．しかし，目録と記述とをどう関連付けしていくかが課題となっている．

この点に関して所内では，左図の移管票へ移管者の側で情報をできる限り多く盛り込んでもらい，これを概要記述に代えるという案や，概要と資料目録とで 2 層化したデータベース（資料群によっては各種メディアの 3 層目を立てる）案が出されている．

また，国語研のウェブサイトでは様々なデータベースが公開されており，これらとの連携も模索しなければならないと思われる．

他方，今後は電子データの移管の増大が予想され，その管理も重要な課題となっている．既に保管してある電子媒体についても，経年劣化によるデータ

▼移管票：移管資料の概要部分（白紙）

資料庫移管資料の概要

調査ID

移管年月日			
移管量	保存箱	メディア箱	カードケース等
移管者名			
受取者名			

調査・研究名称（一般的に研究所で使われている呼称）	実施期間

正式名称および実施期間	
正式名称	期間

担当部署および担当期間	
担当部署名	期間

担当者名

研究・調査の概要（成果刊行物・関連調査等があれば記載）

保存・利用にあたっての留意事項（資料の媒体・個人情報の扱いなど）

▼移管票：移管資料リスト（記入例）

〈資料移管リスト記入例〉

1 紙資料の場合

保存箱番号	内容	所在
1	中学生漢字検証調査 漢字特別アスト(S40.12) 答案(問題1~3)	
2	中学生漢字検証調査 漢字特別アスト(S40.12) 答案(問題4~6)	

2 視聴覚メディア・電子メディアの場合

メディア箱番号	メディア種別	内容	数量	備考	所在
1	オープンリール	「話しことば研究室」録音資料（一般の部） No.1~10	10本		
1	DAT	「話しことば研究室」録音資料（一般の部） 話研1~話研10	10本	同箱のオープンリールをDAT化したもの	
1	MO	「話しことば研究室」録音資料（一般の部） 内容リスト	1点		

↑ 一つの箱の中に，複数のメディアが混在している場合，メディア種別ごとに行を変えるが，箱は一つなので，メディア箱通し番号は同じ

3 カード・その他の場合

カードケース等番号	形態	内容	数量・単位	備考	所在
1	カード	LAJ 原カード	(カードケース幅:¥3×5)×100台		
2	地図	LAJ 調査地点確定用地図	X枚	マップケースでの保管が必要	

の消失が避けられない.

国語研の財産というべき様々な日本語研究資料やデータに関しては，オープン化されたアーカイブをいかに構築するかという課題を残したままであった.資料情報と紙・メディアとを相互にリンクさせ，言語学さらには他領域に開かれた形で言葉の資料を使い・遺していくためのアーカイブが必要となっているのである.

参考文献

1) 国立国語研究所要覧 2014-2015（2014）.
2) 国立国語研究所年報,No.1-45（1949-1994）.
3) 国立国語研究所研究活動一覧（2000-2006）.
4) 青山英幸,記録から記録史料へ：アーカイバル・コントロール論序説（2002）
5) 五島敏芳：アーカイブズ情報の電子化とネットワーク--電子的検索手段の国際規格, 同上（2003）.
6) 森本祥子:アーカイブズの編成と記述--国際動向を中心に，同上（2003）.
7) 山崎圭:アーカイブズの編成と記述--近世史料を中心に,同上（2003）.
8) 森本祥子：EAD を用いた資料記述システムの開発について--国立国語研究所の事例，アーカイブズ学研究，No.4（2006）.
9) 大城博光：公文書目録情報のデータベースモデル--階層構造を持つ目録情報のリレーショナルデータベースでの実装，沖縄県公文書館研究紀要，No.2（2000）.