国立国語研究所所蔵映像資料のデジタル化と所蔵映像データベース構築

石本 祐一（国立国語研究所）
生永 匠（東京電機大学/国立国語研究所）

国立国語研究所には方言，語彙，言語生活，日本語教育，言語コーパスといった，70 年にわたる日本語に関する様々な調査研究による映像資料が保存されている．この映像資料の継承と新たな研究への再利用のために，デジタル化による保存媒体変換と Web アプリケーションによる簡便な視聴環境の構築を行った．本稿では，主な映像資料の紹介とデジタル化の方針，および，視聴環境を提供する「所蔵映像データベース」について報告する．

Digitization of Video Resources in National Institute for Japanese Language and Linguistics and Development of Database for Searching and Previewing Video Resources

Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)
Takumi Ikinaga (Tokyo Denki University / National Institute for Japanese Language and Linguistics)

National Institute for Japanese Language and Linguistics (NINJAL) has many video resources created by various projects that have carried out Japanese language researches such as investigations of dialects, vocabularies, language life, corpora, and so on over the past 70 years. For conservation and reuse of the video resources, the NINJAL have been digitizing them. In this paper, we introduce the video resources and report a policy for the digitization. Also, we show "NINJAL Media Resources Collection" constructed as a web-based system for quickly previewing the videos.

１．まえがき

国立国語研究所（以下，国語研）は，1948（昭和 23）年 12 月の創設以来およそ 70 年にわたって，日本語に関する様々な調査研究を行ってきた．その対象は，方言，語彙，言語生活，日本語教育，コーパスなどの多岐にわたり，それぞれの研究成果は論文や報告書などの形式で刊行されるとともに，近年では国語研学術情報リポジトリ（https://repository.ninjal.ac.jp）としてまとめられ公開されている．また，研究の一次資料に相当する調査票・録音音声・映像・語彙調査の雑誌原本，調査研究運営の記録である調査計画書や会議録，さらに研究成果の前段階である中間生成物に相当する情報カードや集計表なども大量に残されており，国語研研究資料室に保管されている[1-3].

言語資料と言えばかつては文字資料として保存・継承されるものであり，初期の国語研研究資料室収蔵資料も紙媒体がほとんどであるため，資料が眼前にあれば内容の確認ができるという点で参照が比較的容易であった．しかし，近代以降の録音・録画技術の発達によって音声や映像の精緻な記録が可能となり，収蔵資料にも音声や映像そのものが含まれるようになったことで，内容確認には再生機器が必要となった．しかも，録音・録画技術の変遷に伴い記録媒体が移り変わったために，オープンリールテープ，カセットテープ，16mm フィルム，8mm フィルム，VHS テープなど様々な媒体が混在することになり，対応再生機器の旧式化と相まって資料内容の確認を年々困難にしている．

そこで国語研研究資料室において，まずは音源資料の保存・継承を目的として，音源を前代の記録媒体から次代のデジタル記録媒体に移し替える作業を開始した．さらに，音源資料のデジタル化により PC 上での再生が可能となったことを利用して，「所蔵音源データベース」を構築してネットワーク経由で音源の試聴が可能な環境を整えた [4]．すべての所蔵音源資料のデジタル化はまだ完了していないものの，デジタル化に関わる音声フォーマットが定まったことにより容易に増補できる状況である．

一方，音源資料と同様に様々な記録媒体で保管されている映像資料については，貴重な映像が多数存在するものの，保存や視聴のためのデジタル映像フォーマットが定まっていないことからデ

ジタル化が進んでおらず，視聴が困難な状態が続いていた．本報告では，国語研研究資料室収蔵映像資料を紹介するとともに，デジタル化の現状と映像資料の試視聴への対応を考慮した所蔵映像データベースの構築状況について述べる．

２．研究資料室収蔵映像資料

国語研研究資料室で保存している研究資料は，「資料群（フォンド）」と呼ばれる単位で管理されており，概ね 1 課題のもとで生産された研究資料を 1 資料群としてまとめている．なお，2018 年 10 月時点で資料群は 245 である．資料群の概要は「国立国語研究所研究資料室収蔵資料」Web サイト（https://rmr.ninjal.ac.jp/）として 2017 年 3 月より一般公開している．また，個別の収蔵資料に対して，次のような各種目録を整備している．

- 保存箱目録：文書資料や情報カード
- 雑誌目録：　語彙調査に用いた雑誌原本
- 地図目録：　言語地図
- 音源映像資料目録：オープンリールやカセットテープなどの記録媒体

これらは原則として公開されており，来館による閲覧利用が可能である．

以下，映像資料を含む主な研究課題を紹介する．

■ fo0012 談話行動の実験社会言語学的研究

1976–78 年に東京と大阪で座談場面を録画．言語的・非言語的な観点から，言語行動様式を分析することを目的とした調査研究．

報告書：『談話行動の諸相：座談資料の分析』1987 年．

収蔵映像：オープンリール（ビデオ）65 本，VHS ビデオ 3 本，β ビデオ 26 本，U マチック 28 本，DV-CAM 64 本，8mm フィルム 75 本

■ fo0021 調音運動の実験音声学的研究

発音時の音声器官の運動を，X 線映画撮影，ソナグラフ，ダイミック・パラトグラフなどを用いて分析した研究．1965 年と 1967 年に撮影した X 線映画「日本語の発音」は国語研サイトで公開（https://mmsrv.ninjal.ac.jp/x-sen/）．

報告書：『X 線映画資料による母音の発音の研究：フォネーム研究序説』1978 年，『日本語の母音，子音，音節：調音運動の実験音声学的研究』1990 年．

収蔵映像：オープンリール（ビデオ）6 本，β ビデオ 3 本，DV-CAM 72 本，16mm フィルム 69 本

■ fo0025 就学前児童の言語能力に関する全国調査

1965–73 年に実施した幼児の言語能力の習得についての全国調査の一部を収録．

報告書：『幼児の読み書き能力』1972 年，『幼児の文法能力』1977 年，『幼児の語彙能力』1980 年．

収蔵映像：オープンリール（ビデオ）6 本，DV-CAM 6 本

■ fo0059 国際社会における日本語の総合的研究

海外在住の日本人及び日本在住の外国人を対象に，ビデオ刺激による言語行動意識調査を 1996–98 年に実施．

収蔵映像：VHS ビデオ 66 本，8mm ビデオ 104 本

■ fo0074 日本語教育長期専門研修の実施

国立国語研究所日本語教育センターは，1976–2000 年に日本語教師養成のため，日本語教育長期専門研修を実施した．ここで行われたモデルクラスの授業が収録・保存されている．

収蔵映像：β ビデオ 153 本，U マチック 22 本，8mm ビデオ 1 本

■ fo0102 テレビ放送の用語調査

テレビ放送で使用される語彙の調査．1989 年 4 月–1990 年 3 月放送の全国放送網 6 放送局 7 チャンネルの番組からサンプリングした．

報告書：『テレビ放送の語彙調査 1 方法・標本一覧・分析』1995 年，『テレビ放送の語彙調査 2 語彙表』1997 年，『テレビ放送の語彙調査 3 計量的分析』1999 年．

収蔵映像：VHS ビデオ 321 本

■ fo0131 年少者に対する日本語教育の国際的研究

児童生徒に対する日本語教育のカリキュラム開発のため，1995–99 年にかけて意識調査・聴読解調査を実施．調査の一部が収録・保存されている．

収蔵映像：デジタルビデオ 16 本，VHS ビデオ 3 本

■ fo0170 日本語学習者による日本語と母語発話の対照言語データベース

2002–2004 年にかけて中国語・韓国語・タイ語・日本語を母語とする日本語学習者から日本語発話と母語発話を収録．文字化の一部を「日本語学習者による日本語発話と，母語発話との対照データベース」として公開中（https://db3.ninjal.ac.jp/contr-db/）．

収蔵映像：デジタルビデオ 107 本

■ **fo0204 日本語学習者会話データベース（横断調査）**
2006–2009 年にかけて日本語学習者と日本語母語話者との会話を収録．文字化と音声の一部を「日本語学習者会話データベース（横断調査）」として公開中（https://db3.ninjal.ac.jp/judan_db/）．
収蔵映像：デジタルビデオ 44 本

■ **fo0209 話し言葉コーパスの言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築（CSJ）**
1999–2003 年に自発的な「話し言葉」の情報処理技術の基盤確立を目的とした研究・コーパス開発．学会講演等の録画が保存されている．「日本語話し言葉コーパス」（https://pj.ninjal.ac.jp/corpus_center/csj/）は，現在コーパス検索システム「中納言」で利用することもできる．
報告書：『日本語話し言葉コーパスの構築法』2006 年．
収蔵映像：デジタルビデオ 1488 本

このように，一般個人を対象とした自然談話や当時放映されていたテレビ番組もあれば，日本語学習者向けの教育映像や発声中の音声生成器官の X 線映像など多岐にわたる．例えば，1970 年代当時の自然会話は当然現在から遡って収録することはできないものであるし，発声時の X 線映像は現在では倫理的に収録が困難であるなど，いずれも貴重な映像資料である．

３．映像資料の媒体変換・デジタル化

前節の研究課題の紹介で触れたように，国語研研究資料室に保存されている映像資料の記録媒体は研究課題の実施時期および収録方針により様々である．表 1 に保管されている映像記録媒体の種類および点数を示す．このような複数種類の記録媒体で保存していることの問題点として次の二つが挙げられる．

第一は，再生機器の確保の問題である．それぞれの記録媒体に互換性がないことから，映像再生のために記録媒体ごとに対応する再生機器を所持し続けなければならない．さらには，8mm ビデオや DV，U マチック，ベータマックスといった規格の再生機器は生産が終了しており，現有機器の稼働状態によっては映像資料の再生が困難となる．

第二は，記録媒体の保管の問題である．テープ媒体は経年変化によって損傷したり，品質劣化が生じる可能性がある．長期間の保存では適切なメンテナンスを定期的に行うことが必要であるが，

表 1 国語研研究資料室に収蔵されている映像記録媒体

媒体	点数
16mm フィルム	69
8mm フィルム	89
8mm ビデオ	149
デジタルビデオ(DV)	1,761
DVCAM	148
オープンリールテープ	84
U マチック	119
VHS	596
ベータマックス	203
合計	3,218

媒体点数の多さから手入れをすることが難しく，テープにカビが発生して再生が不可能になることもある．

そこで，映像資料の確実な保管を目的として，媒体変換を行うこととした．また，媒体変換と同時に，新たな調査研究のための再利用を容易にすることも考慮して，デジタル映像フォーマットの検討を行った．映像フォーマットは，保存・継承を目的とし可能な限り品質の低下が生じない形式（保存用）と，軽量かつ幅広い環境で再生ができて映像内容の確認が行いやすい形式（確認用）の二種類を用いることとし，次のように定めた．

■ 保存用フォーマット
- AVI 形式
  - コーデック　MDVC
  - アスペクト比　4:3
  - ピクセルサイズ　720×480（または 640×480）
  - ビットレート　約 8.3Mbps

■ 確認用フォーマット
- MP4 形式
  - 映像コーデック　H.264
  - 音声コーデック　MPEG-4 AAC-LC
  - アスペクト比　4:3
  - ピクセルサイズ　640×480
  - ビットレート　約 2.5Mbps

これにより，確認用フォーマットの映像で速やかに内容確認を行い，研究に利用する場合にはより品質の良い保存用フォーマットの映像を使用するという運用を行うことができる．なお，上述の映像フォーマットは現時点の技術を基にした設定であり，将来の映像保存技術の発展に応じた見直しも検討している．

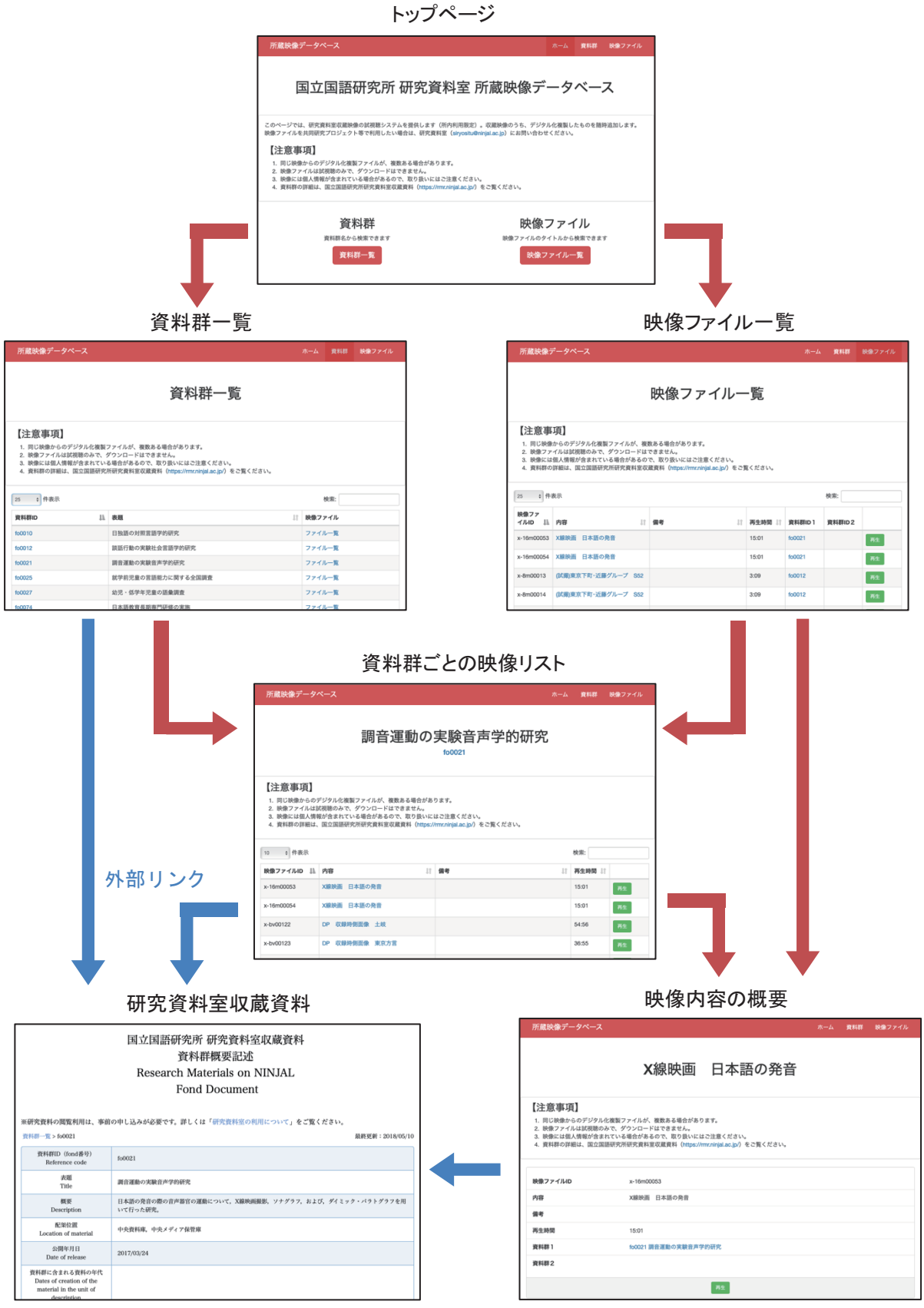

図 1 所蔵映像データベースの画面および遷移図

表 1　資料群テーブル

フィールド	データ型	データ例
資料群 ID	char(6)	fo0012
表題	varchar(1024)	談話行動の実験社会言語学的研究

表 2　映像テーブル

フィールド	データ型	データ例
映像 ID	varchar(64)	x-8m00083
映像名	varchar(256)	文楽(談話行動・身振り)
備考	text	2:50 から映像途切れ
資料群 ID 1	char(6)	fo0061
資料群 ID 2	char(6)	
時間長	char(8)	3:24
ファイルサイズ	bigint(20)	65580419

映像資料はそれぞれの形式でデジタル化したのちに，複数の HDD を組み合わせた RAID 5 構成のファイルサーバに保存することとした．さらに，このファイルサーバのバックアップを定期的に実施することで，データの耐久性を確保している．

4．所蔵映像データベースの構築

音声や映像のような視聴覚資料の研究利用を促すためには，目録提供だけでは不十分であり，どのような内容の音声・映像なのかを研究者が簡便に視聴できる環境が必要である．そこで，国語研研究資料室では，最初に音源資料の試聴に目標を定め，デジタル変換後の音源資料の視聴環境を提供する「所蔵音源データベース」の構築に取り組んだ[4]．所蔵音源データベースは Web アプリケーションとして 2017 年 3 月から運用しており，国語研所外からの接続はできないが，来館者の利用は可能である．先の報告時から大幅に増補し，2018 年 10 月現在において 54 資料群，20,957 ファイルの音源を配信している．

それに加えて，前述の映像資料のデジタル化を基に，映像資料の視聴環境を提供する「所蔵映像データベース」（図 1）を新たに構築し，2018 年 2 月より運用を開始した．所蔵映像データベースは，先行運用している所蔵音源データベースの操作性を踏襲し，各資料群に含まれている映像資料の一覧や個々の映像ファイルのデータ概要を見ることできるようになっている．さらに，資料群の表題や映像名などを対象とした単語検索，ID・

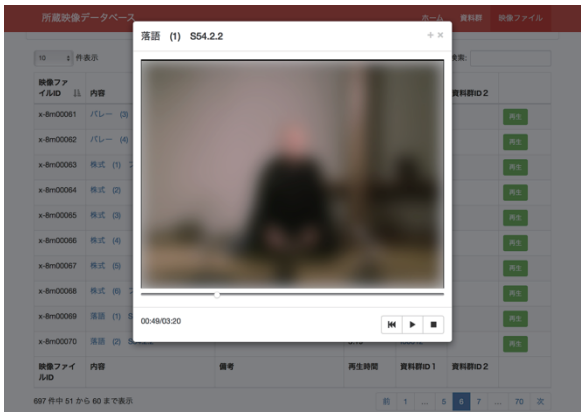

図 2 映像再生プレイヤー

表題・映像名・時間長などによる並べ替え表示もできる．また，同じ資料情報を別々のサイトで個別管理することを避けるために，資料群情報の詳細については「国立国語研究所研究資料室収蔵資料」サイトを参照できるようリンクによる導線を用意するだけにとどめ，所蔵映像データベースでは最小限の資料情報の保持で済むようにつとめている．具体的には，表 1 に示す資料群テーブルと表 2 に示す映像テーブル，そして対応する映像ファイルのみがサイト内に置かれる資料情報となっている．映像テーブルには映像の長さやファイルサイズをあらかじめ保存しておくことでシステム上での情報表示を高速化した．

映像資料の再生は，図 1 中の「映像ファイル一覧」「資料群ごとの映像リスト」「映像内容の概要」の画面に配置された再生ボタン押下により Web ブラウザ上で行われる（図 2）．このとき再生される映像ファイルは前節で述べた軽量な確認用フォーマットであり，ネットワーク経由であっても遅滞なく再生される．

また，所蔵映像データベースは Web アプリケーション上の専用映像再生プレイヤーでのみ再生が行えるように設計した．システムのセキュリティ保護のため技術的詳細は本報告では避けるが HTML5・PHP・JavaScript の組み合わせからなるプログラムによって，利用者が映像ファイルに直接アクセスせずローカルにキャッシュも残さないようにすることで，ファイルダウンロードをさせない仕組みを実現している．これは，所蔵映像資料のなかには個人情報が含まれるものや研究目的でのみ利用可能なものがあることから，ダウンロードによる資料の流出を防ぐためである．このようなセキュリティ対策が導入されているものの，冒頭からの再生だけではなく指定時刻からの再生もほとんど待たずに行えるようになっており，映像再生プレイヤーは資料内容の確認に十分な性能を有している．

所蔵映像データベースは，2018 年 10 月現在，17 資料群の 697 ファイルを配信している．うち，1 時間以上の長さを持つ映像が 3 割程度存在するため，本システムによる視聴環境がなければ内容確認は困難であろう．なお，内容確認後に映像資料の研究利用を希望する場合は，利用申請の上，研究資料室から映像ファイルを提供することになる．詳細は国語研研究資料室の Web ページ（https://www.ninjal.ac.jp/info/aboutus/material-room/）を参照されたい．

５．あとがき

本報告では，国立国語研究所研究資料室の所蔵映像資料と，その内容を Web 上で確認できる「所蔵映像データベース」構築の取り組みについて述べた．これまでは，日本語学・日本語教育の研究者であっても，国語研に音源・映像資料が存在することがあまり知られておらず，知っている研究者であっても実際の音源・映像を確認することが困難であった．先に報告した「所蔵音源データベース」と本報告の「所蔵映像データベース」により国語研内外の研究者に国語研研究資料室収蔵の音源・映像資料が活用されることを期待する．

今後の課題としては，まだ整備が完了していない音源・映像資料も多数存在するため，早期一般公開に努めたい．また，主に構築時期の違いにより音源データベースと映像データベースの 2 種類のシステムに分かれているが，これらを統合して視聴覚資料の総合データベースへと発展させることも検討する必要がある．

謝辞

本研究は，国立国語研究所情報発信プロジェクトの成果を基に行われたものである．

参考文献

[1] 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, 2006, No. 4, pp. 92-102.
[2] 寺島宏貴：日本語研究資料の整備と公開―国立国語研究所研究資料室の取組み，国立国語研究所論集, 2016, No. 10, pp. 245-263.
[3] 山口亮，関川雅彦：国立国語研究所所蔵資料アクセス環境改善への取り組み，人文科学とコンピュータシンポジウム論文集 人文学情報の継承と進化―ビックデータとオープンデータの潮流の中で，情報処理学会シンポジウムシリーズ Vol. 2016, 2016, No. 2, pp. 51-56.
[4] 高田智和，大石恵輔，山口亮，石本祐一：国立国語研究所収蔵音源資料と所蔵音源データベース構築，人文科学とコンピュータシンポジウム論文集 人文学の継承と革新を促進する情報学，情報処理学会シンポジウムシリーズ Vol. 2017, 2017, No. 2, pp. 259-264.