『日本言語地図』のデータベース化

熊谷康雄

１．はじめに

国立国語研究所編『日本言語地図』全６巻は，日本で最初の全国的な規模で行われた言語地理学的調査に基づく言語地図であり，方言研究における基礎資料として広く用いられてきている。調査は1957年から1965年に行われ，編集・刊行は1966年から1974年にかけて行われた。調査項目数285，調査地点数2400である。調査開始から今年で50年になる。

同じ国立国語研究所の『方言文法全国地図』全６巻（調査：1979-1982年，刊行：1989-2006年，調査項目数267，調査地点数807）は，刊行の過程でコンピュータ化を進め，データはインターネット上に公開されている。一方，一世代前の『日本言語地図』の調査および刊行は，調査資料の整理，編集，作図等，全て手作業で行われた。『日本言語地図』の第３集については，佐藤亮一，澤木幹栄，小林隆，白沢宏枝によりデータ化（国立国語研究所（1986））が行われている。このデータ化は原資料に遡って作業が行われているが，今から約20年前，現在から比べると様々な制約のあるコンピュータ環境の中で行われている。

本発表は，『日本言語地図』全体を対象とし，基礎資料として一層徹底した電子化を行うことを意図したものである。

本データベースは，『日本言語地図』データベースと称し，データベース科研の補助（平成13,14,15,16,17年，『日本言語地図』データベース（研究代表者：熊谷康雄））を得て，国立国語研究所において進めてきている。現在，原資料のおおよそ９割程度まで電子化が進んでいる。データの整備を継続しつつ，順次，項目毎の公開を開始する。公開に関する情報は http://www.kokken.go.jp/lajdb/上に掲載していく。

２．『日本言語地図』の資料保存と電子化

言語地図の編集の元になった原資料は，調査者が記録，報告した総数約54万枚の手書きのカードである。このカードは，唯一，国立国語研究所に保管されている。この原資料を後世に確実に伝えていくと同時に，保存と利用の両面を考え，電子的な手段を用いて複製を作成し，データベース化する。

また，物理的なカードは原本そのものであり，電子化したデータに不審な点が見つ

かった場合などの戻るべき拠り所でもある。今後とも，国立国語研究所において将来に渡って保管される。（なお，原調査票は調査者の手許に保管され，国立国語研究所に集められてはおらず，『日本言語地図』はこのカードを基に編集された。）

『日本言語地図』データベースは，(1) 地図化されなかった調査項目も含め，原資料のカードをすべて画像データ化し，また，(2) 『日本言語地図』として公刊した地図上に示された語形の地理的分布情報は文字コードとしてデータ化し，これらをデータベースとして統合的に公開するものである。

原カードの画像データ化は，原資料の保存対策と合わせ，原資料の閲覧・利用を容易にするという側面を持つ。注記などの地図化されなかった情報，凡例の形に統合される前の個々の語形や併用処理の情報などにも簡単に触れることができるようになり，資料批判，新たな観点からの研究や地図化など，研究の一層の深化も期待できる。コード化された情報は，『日本言語地図』に関する情報検索を容易にするとともに，計量的研究を含め各種の研究における『日本言語地図』の利用に新たな基盤を提供する。

また，印刷公刊された『日本言語地図』所収の各地図は，『日本言語地図』データベースと並行して，国立国語研究所において電子化を進め，現在，全 6 巻すべての地図画像をインターネット上に公開している。刊行された『日本言語地図』自体が，研究の基礎資料として参照されてきているものであり，その刊行された姿において，参照可能な形としておくことが必要である。

なお，地図画像の作成には『日本言語地図』の縮刷版を用いた。

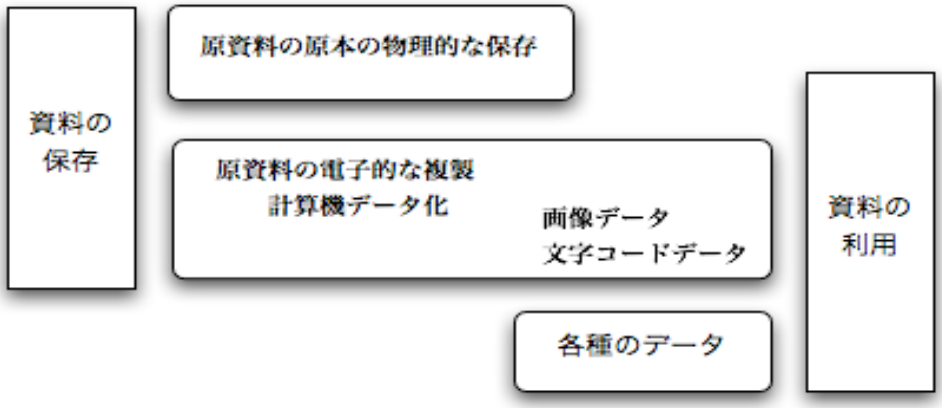


図 1　資料の保存と利用

３．『日本言語地図』の原資料の状態と電子化作業

電子化に関しては，将来のコンピュータの発達を考え，作業開始時の計算機環境の制約の中での効率性を第一とはせず，むしろ，後世に日本の方言に関する基礎資料を残すということを第一義に考え，原資料の複製として，十分にいい品質で整備し，電子化したデータはデータベース化，公開することとして，データの仕様や作業方針等を決めた。

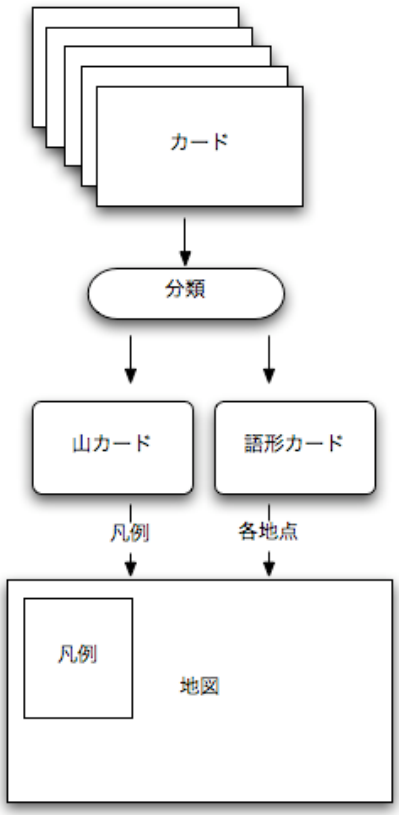
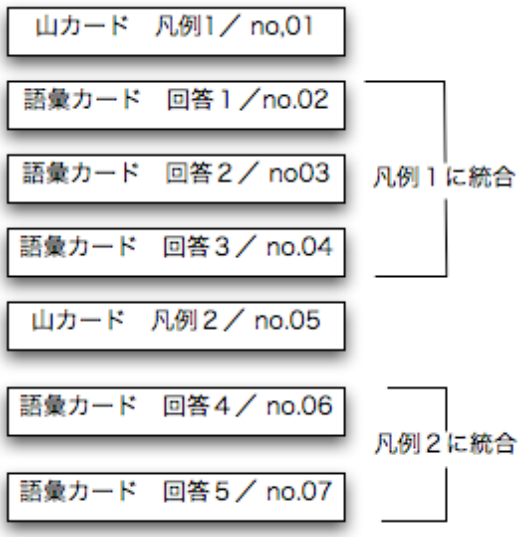


図２　原カードから言語地図へ



語彙カードを分類するため，複数回答が記入されているカードには，その複数回答のパターンに対して，一枚の山カードが作られている。

図３　分類されたカードの配列

現在，『日本言語地図』の原カードは，原則的には項目別に地点番号順に並べられてカードボックスに保管されている。編集作業終了時，地図化のためのカードの分類が行われた配列状態では，カードボックス内では，図３のように，凡例の見出し語形を示す山カードで分類されて個々のカードが並んでいる状態にある。この状態が，図２のように地図上に移される。現在，カードボックスに保管されている原カードは地点番号順に並べられているが，この原カードには，図３の段階の配列の状態でカードの裏に通し番号が振られており，編集作業が終了し，地図化したときのカードボック

スの状態を情報として保存するような仕組みになっている。これは，地図編集当時の所員が，地図化した段階の状態を保存し，編集時のカードの並び順を復元できるように，番号をカードに押したものである。

実際には，これに，地図化の種類（総合図や，語形の前要素，後要素など）に対応する番号も併せて振られているが，基本は，カードボックス内のカードの並び順を保存するための連続番号である。

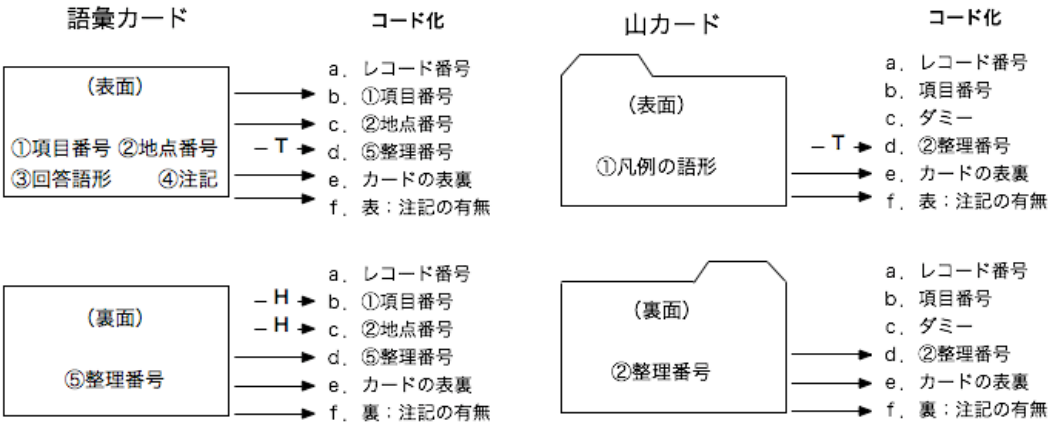


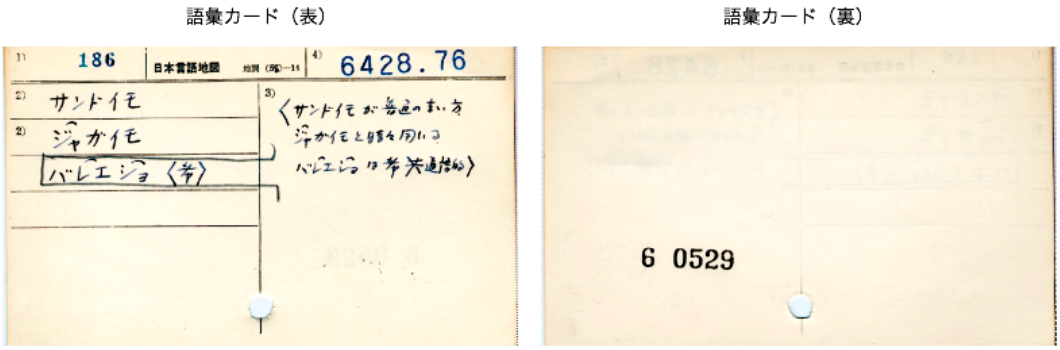
図４　カード上の記載と文字コード化する情報の関係

この番号付け（整理番号と呼ぶことにする）により，カードボックス内のカードと公刊された『日本言語地図』上の分布情報との対応づけが可能となる。

この原カードからコード化できる情報としては，語彙カード上に項目番号，地点番号，整理番号が振られている。山カードには，整理番号が振られている。そして，この整理番号によって，カードボックス内の並び順を復元できるということは，上記のカード毎のコードをデータ化し，この連続番号でソート（実際には逆順にソートする）すれば，計算機上に地図化時のカードボックスの状態を復元し，後は，計算機処理で，各カードにそれが分類される先の山カードの情報を付加してやればよい。これで，凡例上の分類された語形とそれに配属される地点がコード情報としてデータ化できることになる。

原カードのスキャンに際しては，画像ファイル名に上記のコード情報を埋め込むことで，スキャン作業時にコード情報の入力を同時に行った。また，このことにより，画像ファイルと対応するコード情報を実体としてのファイルに一体化して作業対象とできるようにした。

以下，図６に示すような工程を経て，情報の付与，データベース化が行われる。



対応する画像ファイル名

	a　b　c1　c2 d1 d2　e f　g
カードの表面	1177.186.6428.76.6.0529.hn.tiff
カードの裏面	1177.186.6428.76.6.0529.t0.tiff

a: 通し番号４桁，b: 項目番号３桁，c: 地点番号［c1: 前半４桁，c2:後半２桁］，
d: 整理番号［d1: 前半１桁，d2: 後半４桁］，e: 表裏１桁，f: 注記１桁，g: 拡張子１桁

図５　語彙カードの表と裏の画像

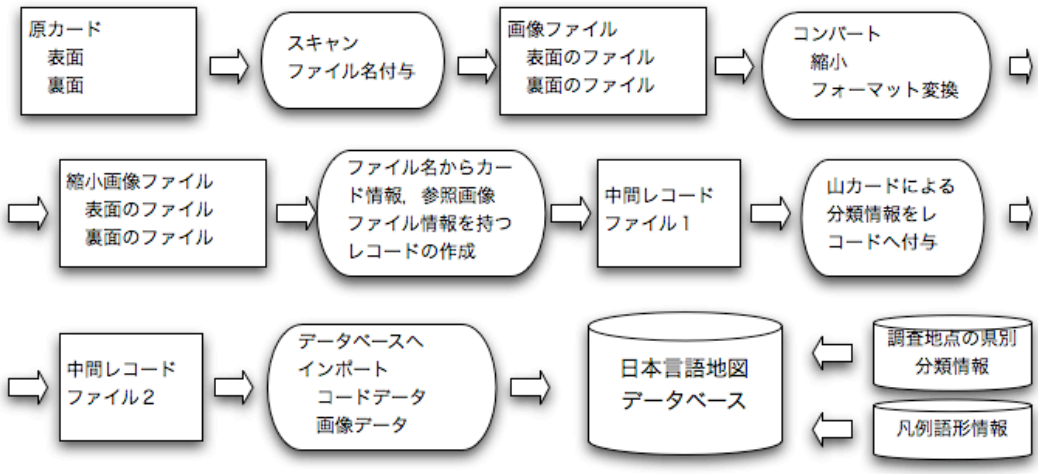


図６　電子化・データベース化の工程（概念図）

カードの配列順を保存する連続番号を振らずに，地点番号順に並べ替えてしまっていたら，上のようにして得られる情報を再現するには，非常に多くの労力を掛ける必要が生じる。仕事の状態を復元できるように番号を振るという判断をしていたことが，今回のデータベース化の計画の実行可能性を高める上で，非常に重要な点であった。後で，

トレース可能となるような情報を埋め込んだということである。現在の立場から振り返ると，非常に重要な判断であった。

４．信頼性の確保と検証手段およびデータ管理と公開

原資料の電子化であるから，そのデータの信頼性（原資料との同一性）の確保が重要である。画像データの持つ情報そのものは，原資料の情報をそのまま伝えるものとして扱えるが，文字コードデータに関しては，誤りが入り込む可能性がある。誤りは，可能な限りゼロに近いことが望まれるが，現実的にはゼロであることを完全に保証することはできない。しかし，正にその地点の情報が重要な意味を持つということはある。

本データベースの場合には，文字コードデータのもとになる原カードの画像と文字コードデータがリンクされており，これが，利用者にもデータの検証手段を与えるものとなっている。地点番号，整理番号は画像ファイルに画像として記録されている。これは，コード化の元の資料にすぐに当れるということである。

また，カードの表と裏で一組であり，１枚のカードに関するコード情報は両面からの情報で構成されている。特に，上のユーザの検証手段を保証するものとして，表裏のペアが正しく組み合わされているかは，重要である。

スキャン作業では，通し番号とタイムスタンプはそれぞれ順序付けられて（原則的に）連続するはずであり，データ化のエラーの回復に役立つ冗長性が組み込まれているということになる。これを利用して，データベース化の際に機械的にエラーを検出し，校正を行う。さらに，同じカードの表裏であるので，ユーザは画像により同じカードの表裏かを見極めることも手段としてはある。

なお，データの整備過程で，あり得ないコードがないか，コードが定義されている範囲内を逸脱していないか，カードの裏表が正しくペアになっているか（通し番号，タイムスタンプの記録）などは，データベース化の際に，機械的なチェックをかける。

公開に際しては，データの維持管理などが重要な意味を持つが，利用法，利用者からの情報のフィードバックも含めて，また，公開したデータのバージョン管理など，基本資料としてのデータの維持管理，情報の共有などを検討を継続しつつ公開を進めていく。その際，データの性格と利用法の段階を考え，公開を加速するため，段階的に公開する方法を検討している。

原データは400dpi，フルカラーでスキャンした（フルカラーのコピー機でカードを原寸大でコ

ピーしたときの品質と遜色のないプリントアウトを得ることができる)。カードの片面8MB，スキャンは両面行うのでカード一枚16MBである。54万枚のカードの表裏両面のスキャンが全て済むと，おおよそ９TBとなる。また，閲覧用に変換した画像ファイルは，平均でおおよそ70KB程度であり，54 万枚のカードの両面で，108万ファイルだとしても，全部で約76GBである。このことは，今のパソコンのハードディスクの中に，『日本言語地図』のカードボックスの全てを入れて持ち運べるということになる。

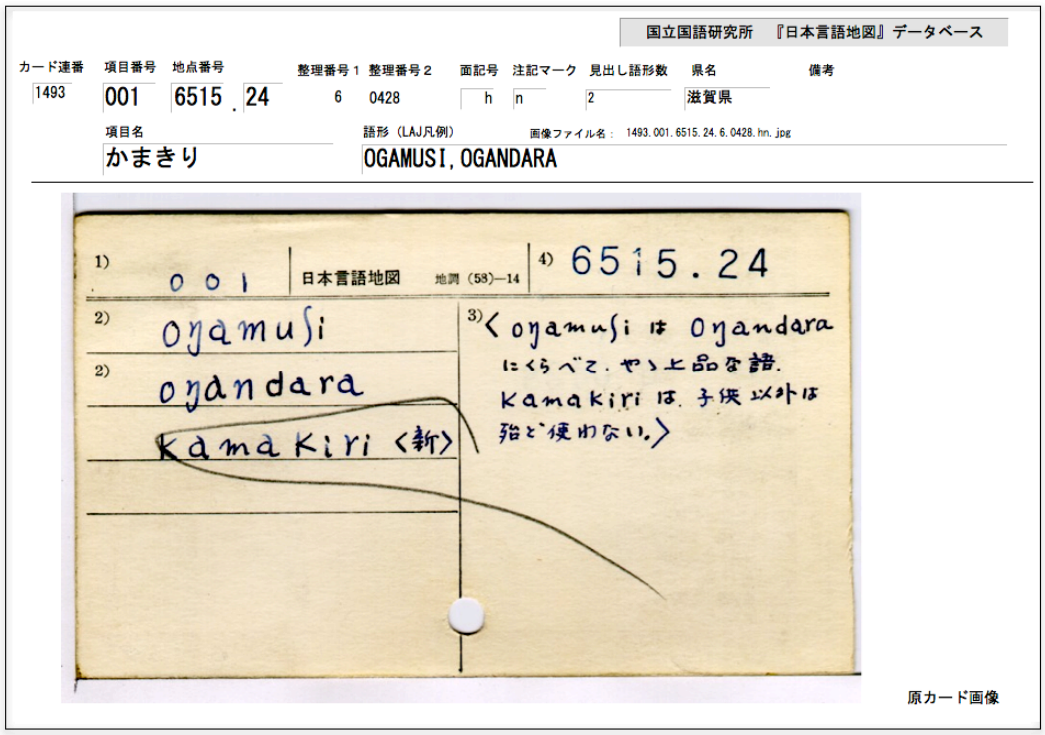


図７　『日本言語地図』データベースの画面サンプル

５．『日本言語地図』データベースの公開

データベースおよびデータは，研究，教育目的には無償で利用可能である。以下のような形で公開する。公開情報は http://www.kokken.go.jp/lajdb/ に掲載していく。

（１）データベース：画面のサンプルを図７に示す。現在の版はファイルメーカを使って作成しており，オンラインおよびダウンロード版を提供する。

（２）コードデータ：コードデータをテキストファイルで提供する。

（３）画像ファイル：原カードの閲覧用画像ファイルを提供する。

６．おわりに

本データベースは平成13,14,15,16,17年度に「『日本言語地図』データベース」（研究代表者：熊谷康雄）として科学研究費研究成果公開促進費（データベース）の補助を受けて行ったものである。原カードに関する『日本言語地図』編集当時の様々な情報は，佐藤亮一，白沢宏枝の両氏に多くを負う。特に作業上遭遇した個々の原カードの不明点に関する情報は白沢宏枝氏から得た。

『日本言語地図』の地図画像の電子化は，国立国語研究所情報資料部門における蓄積資料電子化の一環として行われた。『日本言語地図』のデータベース化も蓄積資料電子化の一環であり，言語資料の蓄積と共有のためのシステムとしてインターネット上に構築している「日本語情報資料館」のコンテンツとしてデータの整備と公開を進めていく予定である。

データベース科研では『日本言語地図』や『方言文法全国地図』に関わるなどの経験を持つ，佐藤亮一，江川清，澤木幹栄，小林隆，白沢宏枝の国立国語研究所 OB の諸氏と，『方言文法全国地図』の編集メンバーである国立国語研究所の大西拓一郎，三井はるみの両氏にメンバーとして協力を得た。『日本言語地図』データベースの計画全体の設計，電子化の実施は熊谷康雄が主として行い，スキャン作業のための原カードの出納管理は礒部よし子が主として担当した。最後に，ご協力いただいた方々に改めて感謝いたします。

文献

国立国語研究所（1966-1974）『日本言語地図』（全６巻）大蔵省印刷局

国立国語研究所（1981-1985）『日本言語地図（縮刷版）』（全６巻）大蔵省印刷局

国立国語研究所（1986）『方言研究と電子計算機』昭和60年度科学研究費補助金（一般研究Ｂ）研究成果報告書（研究代表者：佐藤亮一）

国立国語研究所（1989-2006）『方言文法全国地図』（全６巻）大蔵省印刷局

URL

国立国語研究所「日本語情報資料館」： http://www.kokken.go.jp/siryokan/

国立国語研究所『日本言語地図』データベース： http://www.kokken.go.jp/lajdb/