7. 調査結果データベースの構築

7. 調査結果データベースの構築

7.1. 準備調査結果データベースについて

鑓水 兼貴

①基本方針

「全国方言準備調査」の報告方法は,『全国方言準備調査　調査票』の「手引き」にしたがっている。大西によって作成され，当時の方言調査委員会で了承されたものである。また，それを受けて鑓水によりデータベースの暫定仕様が提案された。基本方針は,

(1) データ入力は調査者が行い，研究所にメールで送信
(2) 書式は「方言文法全国地図 全データ（1～6集)」[1]（以下「GAJデータ」）に準拠
(3) 音声記号は研究所作成のWindows外字ファイル[2]（以下「国研外字」）を使用

としたが，2008年度は，研究所に調査票のコピーを郵送し，研究所でデータ入力のテストをすることになった。音声記号入力用の補助ツールを開発し，データ入力の所用時間の測定などを行った。その後，本調査のデータベース仕様の検討が進むにつれて，音声表記にカタカナを採用するなど，準備調査の暫定仕様から大きく離れたため，2009年度についても，引き続き研究所が入力を担当した。

最終的に準備調査39地点のデータは，すべて国語研究所にて入力を行なった。

②書式

準備調査データベースは,GAJデータを踏襲した。GAJデータの書式は「1行1回答語形」(以下,「回答単位」)を基本とするものである。たとえば，地点Aにおいて○○，△△，××という3語形の併用回答があった場合，

地点	語形
A	○○；△△；××

のように，1行に並べて記述するのではなく，

地点	語形
A	○○
A	△△
A	××

のように，回答数だけ地点Aの行を続ける，というものである。

「回答単位」の書式の利点として，語形の整理がしやすいことが挙げられる。言語地図を作成する場合，語形の整理が重要な作業となるが，回答が行ごとに分かれているため，語形の異なり一覧を容易に作成できる。

---

[1] http://www2.ninjal.ac.jp/hogen/dp/gaj_dp_i/gaj_all/gaj_all.html (2011年2月21日確認)
[2] http://www2.ninjal.ac.jp/hogen/dp/dp_index.html (2011年2月21日確認)

行中に「見出し語形」欄を追加して，

地点	語形	見出し語
A	○○	○○
A	△△	○○
A	××	××

のように，整理した見出し語語形を入れるだけで，整理作業が簡単にできる。

関連して，この書式は GAJ の編集作業のために国立国語研究所が開発した，Adobe Illustrator 用言語地図プラグイン「LMS」[3]での使用に適している。LMS は地点上に地図記号を置く作業を自動化する（併用回答表示も可能）プログラムであり，地点番号とその地点での回答語形（に対応した地図記号の番号）の対応表のファイルが必要であるため，「回答単位」のデータを元に作業をすると，このプログラムを容易に使用可能である。

このほか，行単位のデータ読み込みを得意とする，Awk，Perl 等のスクリプト言語での処理にも適している。

もちろん問題点もある。方言調査では回答に注記情報が付されることが多い。特に，併用回答となった項目に注記がついた場合，「回答単位」の書式では，同じ注記を回答数だけ表示しなければならない。そのため，注記を修正する場合も，回答数ぶん同一内容をコピーする必要があるが，コピーを忘れると回答と注記が分離する恐れがある。

また，地点ごとの行数が固定していないため，回答を一覧しにくいという点も挙げられる。たとえば，

地点	問 1	問 2
A	○○；△△；××	◆◆
B	○○；××	■■■；◇◇
C	○○	■■▼；◆◆

のような「地点×質問項目」の行列の書式は，地点と質問と回答との関係が理解しやすく，データの一覧に優れている。ただし併用回答が多い場合には，セル内が区切り記号などで複雑化するため，データ処理においては，かえって煩雑化する可能性がある。どの書式でも長所と短所があり，「回答単位」と比べてどの書式が優れているということはいえないだろう。

③公開

データベースは，現在（2011 年 2 月 21 日）も細かい修正箇所が残っており，電子版については，当面はプロジェクト内での暫定公開とする予定である。

印刷媒体としては『全国方言準備調査結果データ集』として 2011 年 3 月に公開予定である。

[3] http://www2.ninjal.ac.jp/hogen/dp/gaj_dp_i/gaj_lms.htm（2011 年 2 月 21 日確認）

7.2. 本調査結果データベースの構築

鑓水 兼貴・小西 いずみ・松丸 真大

①XML による「調査票単位」のデータベース（鑓水）

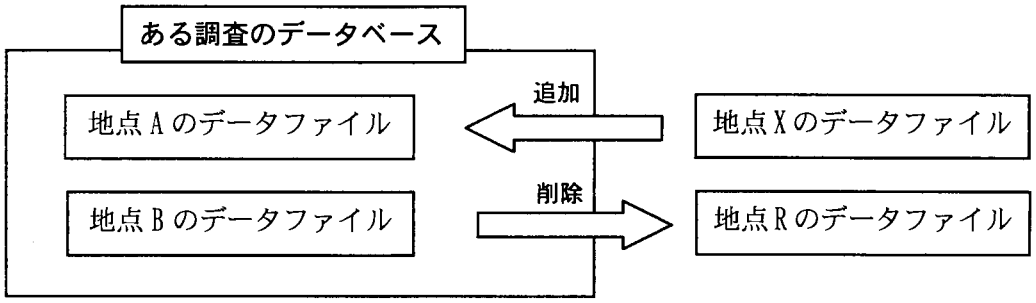
7.1 節でも述べたが，どのデータ書式にも長所と短所があり，ある書式を採用しても，利用に応じて別の書式へ変換する必要が生じる。そのためデータ変換を行うツールの提供は不可欠であり，提供されなければデータ利用の敷居は高いままであろう。

このため，データ変換ツールの提供を前提とすれば，データベースの作成に際して，データの可読性の考慮をする必要はないと判断した。ただし，データベースの保守を考えると，データベースが全くのブラックボックスになることは望ましくない。そのためには，テキストベースで調査報告票の構造を忠実に電子化する書式がよいと考えた。

以上から，準備調査時に計画していた案を変更し，2009 年 12 月に，構造化テキストである XML（Extensible Markup Language）の採用を提案した。

データは「１調査票が１ファイル」(以下,「調査票単位」)となる。データファイルは１話者の調査結果で構成され，調査全体は１つのフォルダに蓄積される。これは，ＰＣでのファイル操作と同じであり，多くの利用者にとってわかりやすいデータベースといえる。

「調査票単位」の利点は，調査時期とは関係なく，自由にデータを追加・削除できる点にある。「全国方言分布調査」は調査中のデータを共有するが，調査票ごとにファイル化されているため，ある地点の調査データが追加されたとしても，残りのデータには全く影響がない。

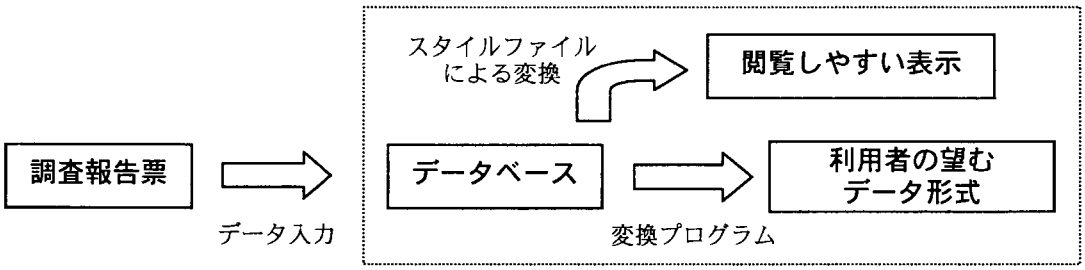


調査票単位のデータベースのイメージ

XML とは，文書中にデータの構造や意味を示す「タグ」を埋め込んで，文書をデータ処理に適した形式にするための記述方法である。情報処理においては広く普及しており，多くのプログラミング言語で XML を処理するためのライブラリが使用可能である。

しかし，XML 文書そのものは閲覧性が低いため，XSLT スタイルシート（表示用のデータ構造変換指定）を用いて閲覧しやすい表示 (HTML など) に自動変換するようにする。

前述のように，XML のデータそのままでは，多くの人にとって利用することが難しいため，利用に際しては「回答単位」など，利用者の望む形式への変換ツールが不可欠である。そのため，データベース単体ではなく，周辺のツール群をあわせた「データベース利用システム」として提供する予定である。



データベース本体に変換等のツールをあわせた
「データベース利用システム」

このことは，他の方言調査データと組み合わせて比較・分析を行う場合にも応用可能である。

データフォーマットを規定すれば XML である必要はないが，多くの調査の場合，質問項目や内容は異なっており，共通する項目を取り出す作業は複雑になってしまう。複数の調査に共通するフォーマットを表形式で作成しようとしても空欄が多くなると，扱いにくいデータになってしまう。

XML であれば，データの内部構造の規定が明示的であるため，調査項目が大きく異なるデータ同士であっても，容易に共通項目を取り出すことができる。これまでの方言調査資料のデータベース化においても，XML 化を進めることで，複数の調査データを組み合わせた研究が可能になると思われる。

こうしたことから，「全国方言分布調査」のデータベースは，他の方言調査資料のデータベース化における指針になりうると思われる。

なお，データベースの概要については，以下の研究会で発表を行った。

2010. 5.25　第 14 回 NINJAL サロン (国立国語研究所)

「全国方言調査におけるデータベース化」(鑓水)

2010.10.30　第 88 回人文科学とコンピュータ研究会 (国立国語研究所)

「方言調査データの XML によるデータベース化」(鑓水・小西・松丸)

2010.12.20　「方言の形成過程解明のための全国方言調査」研究発表会 (国立国語研究所)

「『全国方言分布調査』データベースの概要と利用法」(鑓水)

②語形の表記方法（小西・鑓水）

回答語形の表記は，当初は準備調査と同じ音声記号（国研外字）の予定であった。プロジェクトでは調査期間3年半（2010年度後半～2013年度）で約500地点の調査を予定しており，大規模データを短期間でデータベース化しなければならない。その際に，

(1) 音声記号の入力に時間がかかる
(2) 音声表記の整理・統合に時間がかかる

という2つの問題が考えられた。

調査結果データの入力作業は，外注業者もしくはアルバイトを予定していたが，準備調査データの入力段階で，音声記号の入力に時間がかかることが判明し，特に外注業者の場合には音声記号の知識が全くない可能性が高く，入力には相当の時間がかかることが予想されていた。

また，音声表記の整理・統合も問題となった。音声の微細な表記はデータとしては重要であるが，言語地図の作成など，分析にあたっては一定の基準によって整理しなければならない。GAJにおいても報告された音声表記の整理・統合作業がなされており，これに多くの時間を費やしている。

準備調査では，報告段階で調査者によって音声表記（『全国方言準備調査票』では「日本の方言学で通用しているブロードなIPA表記」と説明）を整理してから提出してもらったが，集まったデータは，この表記から外れたものはわずかであり，ある程度，調査者側で表記の整理を行うことができると考えた。

これらの点を総合して，ワーキンググループの会合において，共通語に近い音素体系を持つ方言に関してはカタカナを用いて表記してもよいのではないか，という意見が出された。カタカナであれば外注業者であっても入力作業が容易である。音声表記としては表現力が低いが，カタカナは言語学に関係しない人々にもわかりやすい表記法であり，成果の公開という点でも望ましいと考えられた。

しかし，カタカナ表記では外来語表記の規則を含めても表現できる音声が少ないことは，方言調査の表記としては問題が残る。そのため若干の補助記号や組み合わせ表記による拡張をする必要があった。ワーキンググループのコーディング担当である小西の主導で，LAJにおけるカタカナ表記なども参考にしながら，原案が作成された。

本土における音韻項目の表記と，琉球方言の全項目については音声記号を用いることとし，文字セットはUnicode(UCS-2)，文字コードはUTF-8とした。

2009年12月のプロジェクトの共同研究者打ち合わせ会（方言調査委員会の後継組織）にて，表記の原案が示された。カタカナ表記への反対意見もあったが，無理にカタカナ表記にするのではなく，音声記号による補足や注記を許すということが説明され，基本的に了承された。

その後，2010 年 3 月に共同研究者の協力でカタカナ表記ならびに新しい報告方法のテストがなされ，問題点を収集し，5 月に改訂案をまとめた。共同研究者の拡大にともない，旧方言調査委員会の枠での議論は終了し，ワーキンググループのメンバーと全国方言分布調査事務局の共同研究者との間で検討が続いた。

その結果，データ入力における省力化の面を重要視し，積極的にカタカナ表記を採用できるように，カタカナ表記の表現範囲を拡大することになった。なるべく音声記号との対応を可能にするため，GAJ において頻出する音声記号について，カタカナ表記への置き換えが検討され，2010 年 6 月に最終的な表記法が確定した。

③語形の報告方法（松丸・鑓水）

語形の報告はおおきく以下の 3 原則を元にしている。

(1) 採否の基準は「話者自身が現在用いる，あるいは過去に使用した語形」である。
(2) 話者自身が現在用いる，あるいは過去に使用した語形を採用語形として報告・記載する。
(3) 話者が使わない語形に関する情報は，注記として報告する。

話者の回答が，使用語形だけで構成されるのであればデータは単純である。しかし実際には回答ごとに付随する情報が得られるため，それらを「注記」としてどう処理するかがデータベース化の課題となる。報告方法の原案作成はデータベース構造担当の松丸主導で行われた。

特に使用しない語形についての注記の記載方法と，それにともなう注記の及ぶ範囲については問題となった。

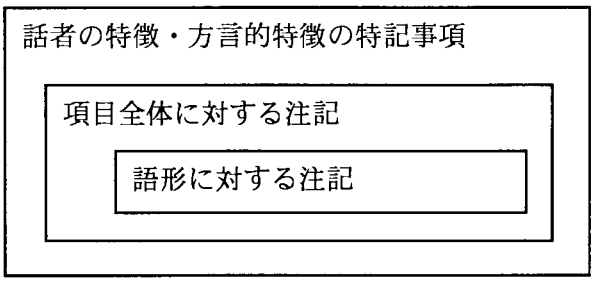
「話者が現在も使う語形」と「話者がかつて使った語形」は，話者自身から両者を区別するかどうかの情報がない限り，特に分ける必要はない。しかし「全国方言分布調査」は過去の方言調査資料との経年比較も主要な研究課題と位置づけられているため，「使用しなくなった」という情報は重要である。そのため「古い言い方で，かつて使ったが現在は使わない語形」という場合には，「〈古〉〈かつて使ったが今は使わない〉」とはせず，新たに「昔」という注記略号を設けた。

同じように，「不使用語形」についての記載方法についても議論がなされた。前述と同様に，過去の資料との経年比較においては，語形の不使用情報は重要である。しかし「不使用語形」の情報を従来のように「語形に対する注記」として扱うと，他の注記の中に埋もれて，不使用語形を取り出しにくくなってしまう。そもそも不使用語形は，ある語形に対しての注記というよりは，採用語形と同じレベルにあるべきで，項目に対する注記として扱うほうが自然と考えた。ほかにも，項目の事物・概念への注記や，項目に関する話者の予想など，項目に対する注記が必要であることがわかった。

以上をふまえ，「語形に対する注記」より上位レベルの注記として，「項目全体に対する注記」を設けることにした。

さらに，注記が複数の項目にまたがったり，調査全体に及ぶような場合の記述方法がな

く，さらに上位の注記として「調査全体に対する注記」の必要性が提案された。当初は，フェイスシートの備考欄での処理が考えられたが，備考欄はあくまでも調査の環境や条件に関する事項であり，言語的な注記は異なる欄を設けるべき，という提案がなされ，「話者の特徴・方言的特徴の特記事項」欄が別に設けられた。このため，注記は，



という三重の構造で記述されることになった。

2009 年 12 月に原案が示され，2010 年 3 月には共同研究者（旧方言調査委員会委員）によって，調査回答の一部を表記案を用いて報告してもらうテストが行なわれた。この結果と，2010 年 5 月の打ち合わせ会議での共同研究者から意見をもとにして，ワーキンググループと事務局の共同研究者によって検討が続けられた。

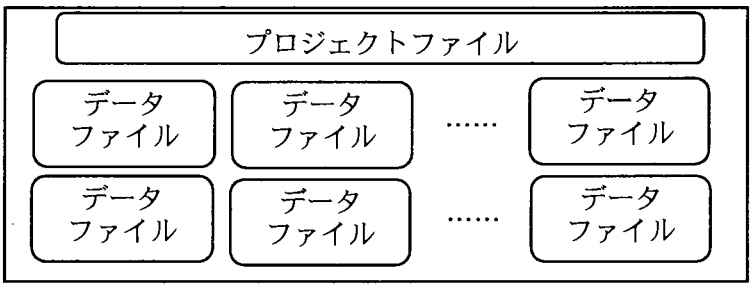
「項目全体に対する注記」は，原案ではデータ構造を意識して回答語形より前に書くことにしていたが，表記テスト時に，注記が語形より前になると，調査の順番と逆になり，直感に合わないという意見が出たため，最終的には全回答語形のあとに「☆」記号を伴って表示することとした。

また，複数の回答に同じ注記が付される場合の表記方法を定めていないことがわかり，「項目全体に対する注記」に準じて，全回答語形のあとに番号を伴って表示することとした。これと同様に，複数の項目にまたがって同じ注記が付される場合についても同様に，調査全体に対する注記ともいうべき「話者の特徴・方言的特徴の特記事項」欄に，「注１」「注２」というように，注番号を付して表示することにした。

以上の流れを経て，2010 年 6 月に最終的な報告方法が完成した。

④方言データベースの構成（鑓水）

一つの調査データは，以下のような構成となる。調査全体の情報を示す１つのプロジェクトファイルと，調査地点分のデータファイルから構成される（図）。以下，プロジェクトファイルとデータファイルの構成について述べる。



方言データベースの構造

プロジェクトファイル

調査同士を比較する上で，調査全体に関する記録をしておく必要がある。単純に比較や集計などができるデータではないが，どのように調査がなされたのかを記録しておくことで，研究者が比較資料として採用するか否かの基準となる重要なものである。

こうした情報が入った調査情報全体を，方言データベースにおいては「プロジェクト」と呼び，以下の３つの情報が入る。これらはXMLによって記述される。

(1) プロジェクト情報
　調査の趣旨・調査概要などを記録する。「話者の選定基準」，「地点の精度」，「語形の採用原則」などの情報が入る。
(2) 調査者情報
　調査者に関する情報が入る。ただし個人情報保護を考慮したデータ化をしなければならない。
(3) 調査票情報
　調査票の情報，すなわち，実際の質問文や，調査時に提示する図などが入る。

データファイル

データファイルは，調査のデータ部分である。以下の３つの部分からなる。

(1) 調査情報
(2) 話者情報
　調査におけるフェイスシートのデータが入る。そのため，個人情報保護を十分に考慮する必要がある。フェイスシートの項目の中には，いつ，誰が，どのような環境で調査したのか，といった調査そのものに関する情報も入っている。これらはXMLタグでは「調査情報」として独立させ，その他の情報については話者情報として，「話者」タグの中に入れる。
(3) 回答情報
　各質問の回答は「調査データ」タグの中に入る。質問文は各個人のデータの中に入れるとサイズが大きくなるため，プロジェクトファイルの中の調査票情報にリンクさせる予定である。

以下に XML のデータ構造の例を示す。(タグ名は分かりやすくするために日本語で表示したが，具体的なタグ名を含むタグセットについては 2010 年度中に公開予定)。

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="chosahyo.xsl"?>
<調査票>
        <調査情報>
                <調査者コード>○○○○</調査者コード>
                <調査開始日>○○○○年○○月○○日</調査開始日>
                <所用時間>○○時間○○分</所用時間>
                <同席者>○○○○</同席者>
                <備考>○○○○</備考>
        </調査情報>

        <話者>
                <話者コード>○○○○</話者コード>
                <話者生年>○○○○</話者生年>
                <話者生育地>○○○○</話者生育地>
                <話者最長居住地>○○○○</話者最長居住地>
                <備考>○○○○○○○○</備考>
        </話者>

        <調査データ>
                <質問項目>
                        <質問番号>○○○</質問番号>
                        <質問項目名>○○○</質問項目名>
                        <質問文>(リンクによって示す)</質問文>

                        <回答>
                                <語形>回答語形１</語形>
                                <注記 種類="話者">語形注記</注記>
                        </回答>

                        <回答>
                                <語形>回答語形２</語形>
                                <注記 種類="話者">語形注記</注記>
                        </回答>
                        <修正 種類="削除" 修正日時="○○○" 修正者="○○">
                                <回答>
                                        <語形>回答語形３</語形>
                                        <注記 種類="話者">語形注記</注記>
                                </回答>
                        </修正>

                        <注記 種類="話者">
                                項目全体に対する注記
                        </注記>
                </質問項目>

                <質問項目>
                    :
                   (略)
                    :
                </質問番号>

                <注記 種類="話者">
                        話者の特徴・方言的特徴の特記事項
                </注記>

        </調査データ>

</調査票>
```

XML によるデータの例

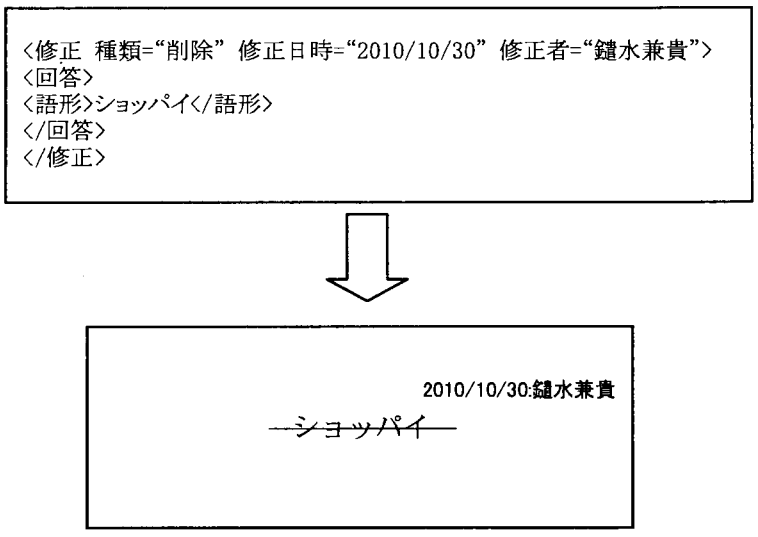
③で述べたように，調査において得られる情報は回答ばかりではない。回答に対する注記，項目に対する注記，調査全体に関わる特記事項などといった，さまざまな補足情報が得られる。これらの情報は，レベルごとに階層化されるため，XML での表記に適している。これらの情報は「注記」タグで表され，提供者が話者か調査者か同席者か，といった情報はタグ属性で分類される。

「全国方言分布調査」データベースは，調査途中から利用するため，データの追加だけでなく，既存のデータ部分にも修正が行なわれる可能性がある。そうした場合，利用者の混乱がないように，リリース後に修正するデータすべて，

〈修正 種類="削除" 修正日時="○○○" 修正者="○○"〉 ●●●● 〈/修正〉
〈修正 種類="挿入" 修正日時="○○○" 修正者="○○"〉 ●●●● 〈/修正〉

のように，「修正」タグを挟むだけで，過去の情報は削除しない。そのため，すべての変更履歴が残される。

「修正」タグは，「削除」と「挿入」からなり，タグ属性によって決まる。また，「修正日時」と「修正者」もタグ属性によって指定され，利用者がどの時点のデータを利用したかがわかるようになっている。これらもスタイルシートによって，修正履歴として表示される。



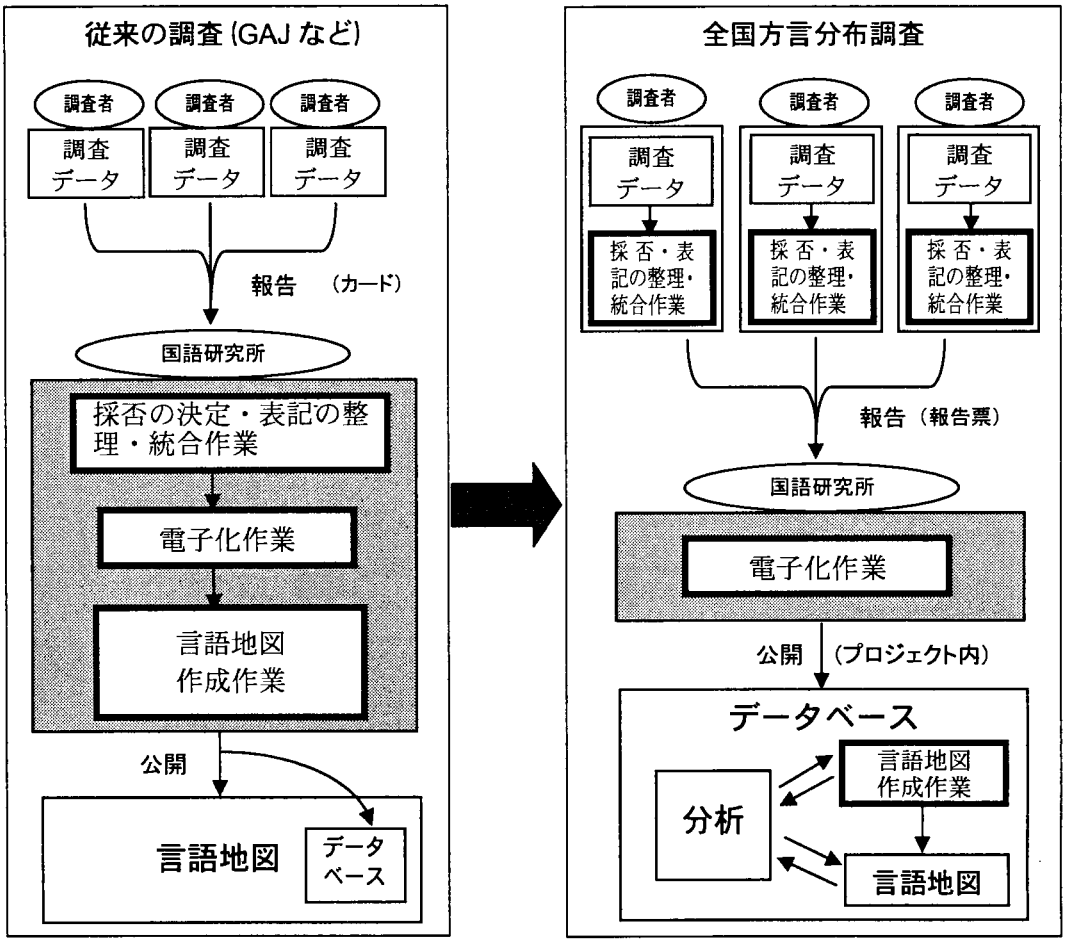
XML による修正タグと表示例

⑤データベース作成までの流れ（鑓水）

下の図は，ＧＡＪなど従来の研究所での方言調査と，本調査とのデータベース作成までの流れの違いについて示したものである。

大きく異なるのは，研究所の役割である。語形の採否や，表記の整理・統合といった作業の多くを調査者側で行ってから，データを提出する。また，言語地図の作成についても，プロジェクト内で公開されたデータをもとに，共同研究者によって作成する予定である。このため，②③で述べたように，表記方法や報告方法について細かいルール作りが行われた。

研究所はデータの入力作業に徹することになり，データ公開までの時間を大きく短縮することで，データの分析に多くの時間を費やすことができる。



データベース作成の流れの比較

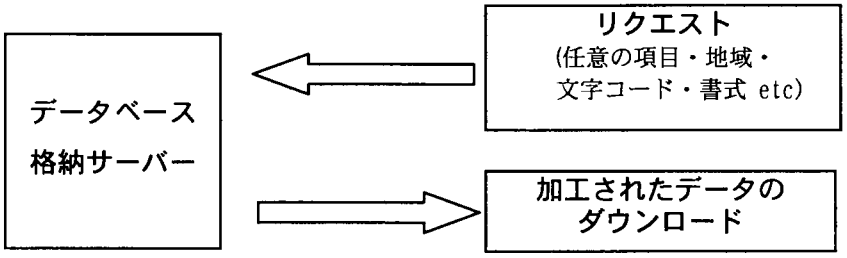
9.3. 方言データベースの意義と研究の展開（鑓水）

方言調査データは，言語地図資料としては大量に蓄積されているものの，データベース化はあまり進んでいない。近年の調査データは基本的に電子化されているが，言語地理学的調査がさかんだった 1970～80 年代の資料は，コンピュータの普及が進んでいなかった（また，性能が低く複雑なデータを扱えなかった）こともあって，電子化は限定的である。多くのデータが現在も紙の資料のままの状態にあるが，これは方言調査では，言語地図を最終報告形式とすることが通例であり，データベースの整備が遅れやすかったことも影響している。

本調査で計画される方言データベースは，この「全国方言分布調査」だけのために設計されたものではなく，そうした他の方言調査データについてもデータベース化の方法として採用されることを目標としている。

7.2 節でも説明したように，本データベースは 1 人の話者の調査結果を 1 つの XML ファイルに記述するものである。基本的には調査票（報告票）の再現を目指しており，調査時の状況を再現しやすくすることが狙いである。また，一部の話者のデータを追加・修正したとしても，残りのデータに全く影響を与えないため，調査期間中でもデータベースを利用しやすい。

XML ファイルのデータは，スタイルシートによる変換によって調査票に近い状態での閲覧が可能となる。さらに，XML ファイルを格納しているサーバー側が，利用者のリクエストした形式に変換して出力するように，データベース利用のためのツールを提供することで，データの分析以外の労力の軽減を目指している。



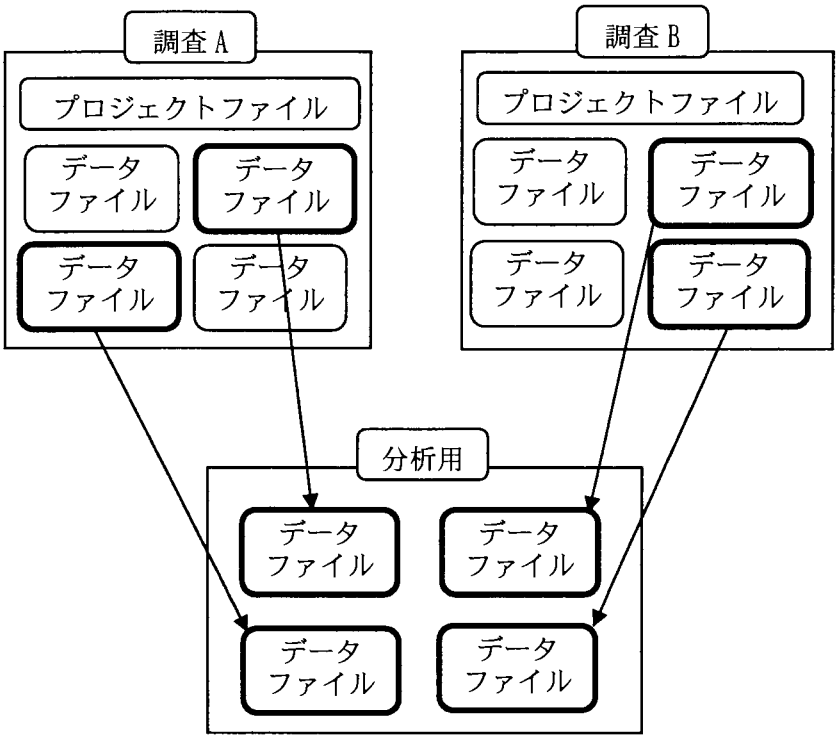
データベースの利用イメージ(1)

データは調査ごとに 1 つのフォルダにまとめられるが，フォルダ内は調査票単位でデータ化されている。そのため複数の調査データを統合するには，データファイルを統合するだけでよい。

たとえば，調査 A と調査 B に共通する地域のデータだけを取り出して新しいデータベースを作成するには，図のように必要な地点の XML ファイルを取り出せばよい。地点は XML ファイル内の属性情報のタグを参照することで，特定の条件に適合した話者を選び出すこ

とができる。両者の調査に共通する「調査項目名」があれば，データファイルの「調査項目名」タグのある項目を取り出すだけでデータの比較が可能となる。

ただし，調査項目の質問の仕方が同じかどうかといった情報は，調査 A と調査 B のプロジェクトファイルを比較して確認しなければならない。



データベースの利用イメージ (2)

また，応用的なことであるが，言語地図作成や計量的分析を支援するツール群の提供も重要と思われる。

方言研究においては，言語地図は分析には必要不可欠なものであるものの，コンピュータを用いた地図作成方法は，今だ普及過程にある状況といえる。また，既存の方言データベースを利用した計量的分析についてもあまり進んでいない。

データベースを提供しても，考察の手助けになるための地図作成や計量的分析の段階でつまづいてしまう恐れがある。そのため，データベースの利用においては，データの加工といった基本的な部分のみならず，言語地図作成のような応用的な部分までを「データベースの利用」として考慮にいれておく必要があるだろう。

以上，XML を用いた方言データベースの利点について述べた。

「全国方言分布調査」のデータベースだけでなく，全国の貴重な方言調査データの共有化を促進することによって，時間的変化・地理的変異を動的に分析するためのデータ整備がなされることが期待される。また，複数の調査を比較するような言語地図の作成や，計

量的分析といった研究も促進されることが望まれる。これらは「全国方言分布調査」における「方言形成過程の解明」という目的にも一致し，今後の方言研究に大きく貢献するであろう。

今後，データベースの利用の過程で出た問題などをもとに，さらにXMLのフォーマットの再検討や改訂を行っていきたい。