『語文』145(2013.3):131-121 (26)-(36)

データ統合・共有を目指したWeb言語地図の構築

—成果公開サイト「日本大学文理学部Web言語地図」の試み—

林　直樹・田中ゆかり

キーワード：WebGIS　Web言語地図　地理的言語データ　データ統合・共有

要　旨

本稿では，異なる研究者によるデータをWeb上で共有・統合することを目的に構築された「日本大学文理学部Web言語地図」の概要を報告する。最初にWeb言語地図の利用方法のうち，言語地図の描画方法を説明する。次に，Web言語地図にデータを追加するために，個人がどのようにデータを管理するのかを述べ，作成したデータをWeb上で管理するための方法を解説する。最後に，Web言語地図の理念である研究資源の共有という試みにおける今後の課題について言及する。

1. はじめに

「日本大学文理学部Web言語地図」（以下，Web言語地図）は，私立大学戦略的研究基盤形成支援事業「東アジアにおける都市形成プロセスの統合的把握とそのデジタル化をめぐる研究」の1グループである，日本語日本文学班の研究の一環として構築された。

従来研究者ごとに管理・分析されていた言語データを一元的にアーカイブする構想に基づき，インターネット上に蓄積したものがWeb言語地図である。この試みは，重要視されながらもなかなか進展しない言語データの共有という理念を，インターネット上において具現化しようという発想に基づく。本プロジェクトが提案する研究資源の共有システムへの賛同者は，参加登録の上，このWeb言語地図にデータの追加・管理が行えるようにした。これは，個別に所有する研究資源の共有化によって，研究資源を有効活用することを意図したものである。さらに，何度でも簡単に言語地図を描画できるようにし，描画しつつ解釈を加えるというような動的なデータ分析も指向した。

研究資源である言語データの共有が重要な課題となることは，すでに荻野（1999a）に指摘がある。一方，荻野（1999b）では，言語研究者の多くが言語データの共有を指向しているものの，アンケートデータを主要な研究資源とするタイプの研究者の間において，言語データの共有化が進んでいないことも示されている[1)]。

言語地図による分析となじみがよい地理的言語データの共有に特化したデータアーカ

イブについても，大西（2002）・出野他（2002）などでその有用性とともに構想が語られているが，いずれも具体的なデータの共有・公開には至っていないようである[2]。

以上のように，1990年代の後半以降，言語データ共有の重要性は研究資源タイプを超えてたびたび言及されてきており，その共有化構想が示された例もあるものの，2013年1月現在において，実現された例はわれわれの確認の限りみとめられない。この背景には，コンピューターの性能や技術的な制約などがあったと思われる。しかし，こんにちではインターネット上でデータをアップロード・ダウンロードするサービスも普及しつつあり[3]，管理者・利用者双方向においてインターネットを介したデータ共有を実現化する素地は整いつつあると考えられる。このような時代的状況の中での新しい試みとして，Web言語地図は構築された。

2. 「日本大学文理学部Web言語地図」概要

Web言語地図は，下記URL（http://www.chs.nihon-u.ac.jp/jp_dpt/nichigo-nichibun/gengo-map/）にアクセスすることで利用できる。本サイトは，タブレット型PCによる閲覧にも対応している。

言語データを描画する地図にはGoogle Mapを採用したため，すべて無料で利用でき，Google Mapの仕様に基づきベースとなる地図の表示を変更することもできる[4]。

Web言語地図の主要な機能は，以下の2つである。

1） 言語地図の描画

2） 言語データの管理

1）は，Web言語地図にアクセスすることによって，誰でも本サイトに格納されているデータを選択した上で，自由に言語地図を描画し，閲覧することができる。一方，2）については，データの信頼性・サイト管理の安全面から，申請の上登録が許可された利用者のみ関与可能なシステムとした。以下では，1）言語地図の描画方法を3.で，2）言語データの管理のために必要なデータの統合・共有方法を4.で，2）言語データの管理方法を5.で説明する。

なお，Web言語地図については日本大学文理学部で開催した展示会「WebGISで体感する江戸・東京」（2012年10月15日～20日），ならびに著者による2012年度日本語学会秋季大会における学会発表（2012年11月4日：林・田中2012）において試験的な公開を行った。それぞれの場において受けたコメントを反映させるかたちで，プロジェクト期間満了時に公開予定の最終公開版[5]ではいくつかの機能を追加する。追加予定の機能については，各章当該部分において順次述べていくことにする。

3. 言語地図の描画

ここでは，Web言語地図を用いた言語地図の描画方法を，実際の画面を示しながら説明する。Web言語地図でWeb上に描画できるのは，次の3種のデータである。

1) 言語情報データ
2) 話者情報データ
3) 鉄道データ

以下，それぞれの描画方法について順に述べていく。

3.1. 言語情報データ

まず，言語情報データの描画方法を示す。Web言語地図は，記号の種類で表されるアイテムデータと，記号の大小で表される数値データという2種類の言語情報データを，それぞれに適した形で地図上に描画できる。

① アイテムデータの描画：画面左上のアイテムボックスのうち，「個別」タブに登録されているデータから任意の項目を選択する。項目を選択すると，当該項目が出現する地点に☆や△などのアイコンがプロットされる（図1）。

図1：言語情報データ描画画面（アイテムデータ）
「厚い」終止形0型を用いた描画

② 数値データの描画：画面左上のアイテムボックスのうち，「出現度数」に登録されているデータから任意の項目を選択する。項目を選択すると，選択した項目の出現度数に応じた円が地図上にプロットされる（図2）。

③ 描画画面の操作：図1・2のように描画された地図は画面右上のバーで拡大・縮小といった操作が行える[6]ため，同一のデータに基づいて広域・狭域の地図を表示することもできる。

④ 描画画面の保存：描画画面を保存する場合は，それぞれのOSに応じた方法で画

面のキャプチャを行う[7]。

⑤ 準備中の追加機能：どのような特徴が何度数現れ，それがどのように表されているのかを表示する「凡例」がある。この機能が実装されれば，アイテムデータと数値データを同時にプロットする場合などでも，何が表示されているのか迷わないようになる。

図2：言語情報データ描画画面（数値データ）
終止形2型の出現度数を用いた描画

3.2. 話者情報データ

言語地図データを描画すると，プロットされた地点の話者がどのような属性であるのかを以下のような形で確認することができる。

① 話者情報の表示：アイコンがプロットされた地点を選択すると，生年や性別など，当該地点の話者情報がポップアップ形式で表示される[8]。

② 絞り込み描画：話者情報データから必要な情報を選択すると，絞り込み描画を行うことができる[9]。この機能により，調査者，調査年，回答者の生年・性などの属性を指定した結果のみを用いた言語地図の描画が可能となる。

3.3. 鉄道駅データ

Web言語地図では，私立大学戦略的研究基盤形成支援事業のプロジェクトを構成する1グループである地理情報班から提供を受けた，首都圏（1都3県）鉄道駅データをWeb言語地図で展開できる。

① 鉄道駅データの描画：画面左下のアイテムボックスのうち，「路線駅」に登録されているデータから任意の項目を選択すると，路線ごとの鉄道駅がプロットされる（図3）。

② 鉄道駅の開通年・廃止年の選択表示：鉄道駅の開通年・廃止年といった情報を選択すると，特定の年代の鉄道敷設状況を参照できるようになる。首都圏の言語事象分布の背景にあると想定される，言語形成期当時の交通アクセス状況を視覚的に検証することが可能となる。

図3：鉄道駅データ描画画面
2013年1月現在の山手線駅を用いた描画

4. 言語データの統合・共有

次に，上記のように言語データを描画し，Web上にアーカイブするための，データの統合・共有方法を解説していく。なお，これらの作業はテンプレートに基づいて行えるよう，テンプレートデータとテンプレートデータの作成マニュアルをダウンロードできる機能をWeb言語地図に搭載している。

ここではWeb言語地図にすでにアップロードされている試行データを例に取り，データの統合・共有プロセスを具体的に示すことで，個人が所有するデータをWeb言語地図上に追加できるようなデータ形式に整える方法を示す。

4.1. 試行データ

ここで説明のために用いる試行データは，首都圏を調査対象とした異なる調査者による2種類の調査データである。ひとつは首都圏西部域データ（調査者：田中ゆかり，調査年：1992年，分析対象者：72人），もう一方は東京東北部データ（調査者：林直樹，調査年：2010年，分析対象者：44人）。どちらも主に高年層（調査時60歳以上）を対象としたリスト読み上げ式の面接調査に基づいている。試行のための分析項目として共有したのはアクセントデータで，3・4拍形容詞4語の終止形と連体形，計8項目である。

4.2. データ形式の統合

以下，データ形式の統合方法について具体的に述べていく。

① ファイル形式の統合：まず，それぞれが管理するオリジナルデータのファイルをCSV形式にする[10]。

② 入力情報の統合：Web言語地図にアップロードするために必要な，(1) 話者情報データ・(2) 言語情報データをそれぞれ用意する[11]。

(1) 話者情報データ：話者情報データは，当該話者の生育年・性別・出身地・緯度経度情報など，フェイス項目を管理するためのデータである（表1）。

表1：データ形式例（話者情報データ）

ID	名前	性	生年（西暦）	生年（年号）	調査年	調査者	調査時年齢	緯度	経度
1	T-1	1	1912	明治45年	1992	田中ゆかり	80	139.48488	35.92164
2	T-2	1	1915	大正 4年	1992	田中ゆかり	77	139.32042	35.86208
3	H-1	1	1946	昭和21年	2010	林　直樹	64	139.79831	35.75768
4	H-2	1	1924	大正13年	2010	林　直樹	86	139.88835	35.69277

これは，話者がどのような属性を持つのかを管理するためのデータといえる。この中には，当該話者情報を地図上にプロットするための地点情報が，緯度経度として入力されている[12]。

(2) 言語情報データ：言語情報データは，当該話者にどのような言語的特徴が現れたのかを管理するためのデータである。このデータは，語彙・文法・音声などの特徴を問わず入力することができる。アイテムデータの場合は，対象とする特徴が出現したか否かを，0＝非出現・1＝出現，というような0／1形式で入力する。数値データの場合は，対象とする特徴が何度数出現したかを連続した数値で入力

する。この形式で整えられた試行データを表2に示す。

表2：データ形式例（言語情報データ）

ID	名前	厚い 終止0	厚い 終止2	厚い 連体0	厚い 連体2	3拍形容詞Ⅰ類 終止2	3拍形容詞Ⅰ類 連体2
1	T-1	1	0	1	0	0	0
2	T-2	0	1	1	0	1	0
3	H-1	0	1	0	1	1	1
4	H-2	1	0	1	0	2	2

表2では，試行データの左から3列目以降にアクセントデータが入力されている。3列目は「厚い」終止形が0型（LHH）で発話されたか否か，4列目は2型（LHL）で発話されたか否かを表している。表中グレーの網掛けで表示した部分は数値データで，3拍形容詞Ⅰ類・終止形，ならびに連体形において2型（LHL）が何度数出現したのかを表している。このようにアイテムデータと数値データを異なる形式で管理することによって，3.1で示した2種のデータタイプ別の描画が可能となる。

③ 話者情報データ・言語情報データをつなぐキー：表1・表2で示した話者情報データと言語情報データから構成される試行データは，それぞれは「ID」と「名前」が共通しており，この項目が2つのデータをつなぐキーとなっている。話者情報データと言語情報データのキーが一致しない場合はデータが正確に構築されず，地図上にプロットすることもできないので注意が必要である。

以上のデータ作成方法は，Web言語地図右下に掲げられている「マニュアル」から閲覧することができ，テンプレートデータもダウンロード可能となっている。

5. 言語データの管理

次に，4.の手続きを経て作成されたデータをWeb言語地図で管理するための方法を解説する。先にも触れた通り，Web言語地図を用いたデータの登録・管理は，申請許可を受けた利用者のみが参加できるシステムとしている。以下では，参加登録，データの管理の方法について示す。

5.1. 参加登録

Web言語地図でデータを管理するための個人画面に進むためには，事前に参加登録を行う必要がある。参加登録を行う際は，日本語日本文学班ポータルサイトに掲載してい

る利用規約・参加規約を閲覧し，それらを理解の上，賛同できるようであれば，Web言語地図画面右下の参加登録画面にて諸事項を記入し，申請手続きに進む。参加登録が許可されると，本サイト管理運営者から，管理画面のURL・ID（メールアドレス）・パスワードが参加登録者の申請したID（メールアドレス）に送付される。

5.2. 管理画面による言語データの管理

① 管理画面へのログイン：本サイト管理運営者から送付されたURLに移動し，ID（メールアドレス）パスワードを用いて個人の管理画面にログインする（図4）。

図4：Web言語地図データ管理画面

② データの追加：管理画面左中ほどの「csvインポート」を選択し，データをアップロード・インポートする（以下，本稿ではデータをWeb上に追加することをアップロードとする）。なお，一度アップロードされたデータは基本的にそのままWeb上に集積される[13]。これは，アップロードされたデータの履歴管理と，すでに投稿されたデータで描画された言語地図の再現性を確保するための措置である。

③ アイテムデータの管理：管理画面左上の「個別」タブを選択し，地図上に配置するアイコンのタイプならびに色を調節する。現在選択できるアイコンは8種類，選択できる色は12種類であるため，最大で96パタンまで対応することが可能である。

④ 数値データの管理：管理画面左上の「出現度数」を選択し，出現度数に応じた円の大きさを調節する。出現度数の範囲も自由に設定できる。たとえば，設定した階層よって円の大きさを変える，といった変更も可能である。

⑤ 変更データの確認：このような手順を踏んで統合されたデータは，3.1の地図画面ですぐに確認することができる。

⑥ データの修正・削除：Web言語地図では，一度投稿したデータを修正する場合は，修正データを新規データとして再度アップロードする必要がある。これは，アップロードされたデータの履歴管理と，すでに投稿されたデータで描画された言語地図の再現性を確保するための措置である。よって，データを修正する場合は自身が所有するオリジナルデータを修正し，改めて②の手順を踏むことになる。

6. Web言語地図を利用する際の注意点

最後に，Web言語地図を利用する際の注意点について，2点ほど簡潔に述べる。

Web言語地図を利用して描画・分析を行う場合，Web言語地図に虚偽のデータや不正確なデータが紛れることがないとはいえない。このような問題を考慮したため，前述したように参加登録がなければデータの管理はできないようなシステムをとっている。しかし，Web言語地図を用いて分析を行う場合は，登録データの性質・属性をよく理解した上で使用する必要があるだろう。データの性質・属性はデータ選択画面で閲覧することができる。

また，Web言語地図の公開データのみを用いた報告などがなされる場合も想定される。本サイトはCreative Commons Licenses（http://creativecommons.jp/licenses/）の考えにのっとり，研究・教育を目的とした非営利利用については，「クレジット表示・非改変」の条件の下，申請不要・無料での二次使用を許可している[14]。そのため，本サイトにおけるデータ共有に同意した提供者本人が意図しない形で二次利用される可能性があることについても，十分な理解が必要だろう。

7. 今後の課題

以上，日本大学Web言語地図を用いたデータ描画方法，Web上にデータをアーカイブするためのデータ統合・共有プロセス，ならびにその管理方法を解説した。Web言語地図を用いたデータの管理・分析が広く行われるようになれば，より動的なデータ管理・分析に繋がると予想される。しかし，本Web言語地図は未だ試行段階にあるため，運用をしながら本システムの改善をしていくことになる。そのためには，本サイトの利用者・登録者がひとりでも多くなることを期待したい。

その他予期される課題として，利用者が増えた場合のデータ形式の不統一，著作権にかかわる問題などがある。

本サイトの利用に際して，お気づきの点などがあれば，国文学科または本プロジェクト専用のアドレス（nichigo-nichibun@chs.nihon-u.ac.jp）まで，ぜひお知らせいただきたい。

【付記】

本研究は，私立大学戦略的研究基盤形成支援事業「東アジアにおける都市形成プロセスの統合的把握とそのデジタル化をめぐる研究」（研究代表：加藤直人）によるものである。

【謝辞】

調査ならびに本プロジェクトにご協力くださった多くの方々に厚く御礼申し上げます。Web言語地図の改訂にあたっては，「WebGISで体感する江戸・東京」展示会（2012年10月15日～20日），ならびに富山大学で開催された2012年度日本語学会秋季大会（2012年11月4日）の場で受けたコメントを参考にしました。鉄道駅GISデータは同プロジェクト地理情報班から提供を受けました。

引用文献

出野憲司・大島一郎・久野マリ子・竹林暁（2002）「インターネット上で方言情報を共有する試み―東京都言語調査から―」『第16回日本音声学全国大会予稿集』,pp.65-70.

大西拓一郎（2002）「全国型資料と調査の課題―JDnet構想―」佐藤亮一・小林隆・大西拓一郎編『方言地理学の課題』明治書院,pp.389-402.

荻野綱男（1999a）「言語研究と言語データの共有 第1回調査の主旨と電子メール使用」『日本語学』18（9），pp.118-122.

荻野綱男（1999b）「言語研究と言語データの共有 第6回データ収集に関する問題」『日本語学』19（2），pp.90-96.

林直樹・田中ゆかり（2012）「地理的言語データの統合的分析―首都圏アクセントを事例とした試行―」『日本語学会2012年度秋季大会予稿集』pp.241-246.

参考サイト

電通総研メディアイノベーション研究部「クラウドサービス利用者への利用実態と意識に関する調査」（http://www.dentsu.co.jp/news/release/2012/pdf/2012103-0920.pdf）（2012年12月26日最終閲覧）

注

1）荻野（1999b）では，電子コーパスを主要な研究資源とする研究者に比べ，アンケートデータならびに実験データを主要な研究資源とする研究者は，データ共有の割合が低いことを示している。

2）出野他（2002）では，『新・東京都言語地図』のデータを用いたサイトを公開する予定でURLが記載されているものの，2012年12月26日現在閲覧することはできないようである。

3）2012年9月に行われた調査によると，ネット上で任意にデータ管理ができるクラウドサービスを推定748万人が利用している（電通総研メディアイノベーション研究部）。

4) 通常の地図・地形地図・航空写真の3種類が選択できる。その他，画面右に表示されている人型のアイコンをドラッグすることにより，Google Earthに移行することもできる。ただし，Google Earthではアイコンなどは配置されない。

5) 2013年3月末日予定。

6) タブレット型PCの場合はピンチイン・ピンチアウト（2本の指を画面上に載せてその間隔を縮める，もしくは広げる動作）を行うと，拡大・縮小が行える。

7) WindowsであればPrtScキーで，Macであれば「コマンド+Shift+3」キーで行うことができる。Web上の画面を画像として保存するサービスは，Web Screenshots（http://ctrlq.org/screenshots/）・Fast Stone Capture（http://www.gigafree.net/tool/capture/faststonecapture.html）などがある。

8) 管理上の理由により，すべてのデータを一覧することはできない。

9) 林・田中（2012）においてWeb言語地図を用いて調査時期・調査地域の異なるふたつのデータを統合の上，言語変化過程の解釈を試みたところ，データを一元的に描画する方式をとっていたために調査時期の違いが反映されず，誤読の可能性をもつことが確認された。異なるデータを一元的に描画する利点も生かしつつ，このような誤読を回避するために，描画対象データの選択機能を追加することとした。

10) ソフトへの依存がない点で汎用性が高く，2013年現在一般的と思われるためにCSV形式を採用した。Excelなどのソフトでデータを作成した場合は，CSVに変換して保存を行うことでアップロード可能な形式となる。「名前を付けて保存」→「ファイルの種類」でCSV形式を選択して保存。

11) 話者情報データ・言語情報データを別々に管理せず，一括管理する場合のデータにも対応できるよう，テンプレートデータとテンプレートデータ作成マニュアルをWeb上に掲げている。

12) 調査地点データを住所で管理している場合は，東京大学空間情報科学研究センターが提供する「CSVアドレスマッチングサービス」（http://newspat.csis.u-tokyo.ac.jp/geocode/modules/addmatch/index.php?content_id=1）を使用することで，緯度経度に変換することが可能である。

13 退会後もWeb言語地図上での公開，提供データの共有化が継続される。詳細は，ポータルサイトから閲覧可能なWeb言語地図の利用規約・参加規約参照。

14) 本サイトのCCライセンス条件は，「権利者表示-非営利-改変禁止（CC-BY-NC-ND）」。保護期間切れまたは権利放棄相当のパブリック・ドメインはもとよりこの限りではない。また，営利目的の二次使用については別に定める利用規約にのっとり事前申請の上，有料と定める。詳細は，ポータルサイトから閲覧可能なWeb言語地図の利用規約参照。

（はやし　なおき，大学院博士後期課程・文理学部RA）
（たなか　ゆかり，本学教授）