データベースシステム 128－35
（２００２．７．１８）

曖昧な地名照合手法を用いた生物種標本の地図ブラウザ構築

相良 毅[†1] 松浦 啓一[†2] 佐藤 聡[†3] 志村純子[†4]

[†1]東京大学空間情報科学研究センター，[†2]国立科学博物館
[†3]筑波大学学術情報処理センター，[†4]国立環境研究所

生物資源の保全と継続的な利用のために，生態系のサンプリングデータである生物種標本のデータベース化が進められている．生物種の分布を知るためには標本データに含まれる場所の情報（生息地，採取地など）から視覚的な分布図を作成する必要がある．そこで本研究では，標本データに含まれる曖昧な場所の記述を解析するための地名辞書（Gazetteer）を作成し，それを用いた地名照合手法の開発を行い，地図化するブラウザを実装した．

Development of a specimen map browser using a robust geo-coding algorithm

Takeshi Sagara[†1], Keiichi Matsuura[†2], Akira Sato[†3] and Junko Shimura[†4]

[†1]Center for Spatial Information Science, at the University of Tokyo, [†2]National Science Museum
[†3]University of Tsukuba, Science Information Processing Center, [†4]National Institute for Environmental Studies

For conservation of ecosystem and sustainable use of biological resources, development of a specimen database is quite important. In order to know distributions of species, it is necessary to create a visual distribution map from place descriptions included in specimen data (a habitat, extraction ground, etc.). Therefore, creation method of the place name dictionary (i.e. gazetteer) for analyzing ambiguous place descriptions included in specimen data and geo-coding technique using it were developed. We also implemented a map browser for the specimen database.

1. はじめに

近年，開発途上国では工業化のための開発が急ピッチで進められているが，これらの地域では生物資源の調査を行う資金・技術の問題から，十分な調査が行われないままこれまであまり人間の手が入っていない地域の希少な生態系が失われている．そこで，生物資源の継続的な利用を目的として国際的に協調して生物の多様性を保全しようという活動が進められており[1]，その中でも重要な活動の一つに生物種のデータベース化がある．特に各国の博物館などの研究機関が保有する生物標本は，生態系を知る上で重要なデータとなる．生物標本には統一された記述様式はないが，一般に種の分類情報や名称，採集された日付や場所などの情報が付されている．これらの情報からはさまざまなデータベースを構築することが可能だが，ここでは種の分布図を作成することを目的とした．種の分布図を作成することには次のような利点がある．

- 環境要因（気候，植生，水質，土壌，標高など）の情報と重ね合わせることにより，種の生育条件を知ることができる
- 希少種の保護など，生態系の保全策を効果的に立案することが可能になる

生物標本の分布図を作成する上で問題となるのは，場所の表記が極めて曖昧であるということである．

最近では GPS を利用して標本の採集地を正確な緯度，経度で記録することが可能になってきたが，既存の標本ではこのような座標値の記載はほとんど行われておらず，「人間が読めば分かる」という曖昧な基準で地名が記述されている．また，標本の採集者が小中学校の教諭やアマチュアの場合や，上述のような開発途上国では GPS による正確な採集地の記載を常に期待することはできない．そのため，曖昧な地名の表記から緯度，経度のような座標値への変換を行う地名照合処理が必要となる．

本研究では，生物標本データベースとして国立科学博物館所蔵の魚類標本を対象として，曖昧な地名の表記を理解し適切な座標値への変換を行う地名照合処理を開発し，変換結果を地図として表示する地図ブラウザを実装した．2 では魚類標本データに含まれる地名記述の特徴と，地名照合に用いる地名辞書を作成する手法を説明する．3 で開発した地図ブラウザを示し，4 でまとめる．

2. 標本データにおける地名記述

2.1. 生物系データベースにおける地名表記

生物系のデータベースは，植物，動物，微生物など，対象とする生物群に合わせて含まれているデータ項目が異なる．例えば脊椎動物は長年にわたる分類と観察が行われており，種の生息場所がある程度分かっているため，生息場所の情報が含まれていることが多い．しかし，微生物のように，条件さえ整えばどこにでも生息する生物群に対しては，生息場所の情報よりもその株が保存されている場所（つまり研究用に分けてもらう場合の問い合わせ先）の情報が含まれていることが多い．このようなばらつきはあるが，文献[2]によれば，日本国内で電子化されている生物系データベースの 41.6%に生息地，生育地などの場所情報が含まれている（電子化されていないデータも含めると 47.7%）．また，場所情報の内容は，写真撮影地，（推定）分布，発見地，標本採集地などの地名である．

2.2 国立科学博物館所蔵 魚類標本データ

本研究で対象としたデータは，国立科学博物館所蔵の魚類標本データベースより抜粋した，日本国内で採集された標本の 28,555 レコードである．また，データ項目として和名，属名，種小名，採集地を取り出した．このデータベースに含まれる地名の例を

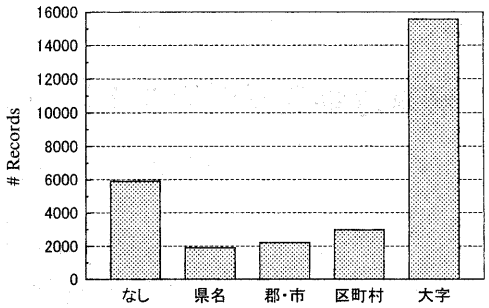

図 1 魚類データベースに含まれる住所の詳細度

示す．

「広島県豊田郡本郷町，沼田川本郷橋」
「利根川本流：茨城県鹿島郡波崎町太田地先」
「埼玉県北葛飾郡幸手町權現堂川」
「北海道，紋別郡雄武町御西川中流」

これらの例のように，採集地の地名は「住所」と「河川名」の組み合わせが圧倒的に多く，全体の約 7 割を占めている．その他の表記としては，住所のみのものや，水産試験場名が記載されているものなどである．住所表記の詳細度は県名だけのものから字名まで含まれているものなどさまざまで，図 1 のように分布している

さて，このデータベースをより正確に地図化することを考える．採集地の地名を一般的な地名照合によって座標値に変換すれば，県名以上の地名を含むレコードをとりあえず地図上にマッピングすることができるが，河川名の情報が利用できない．例えば「京都府由良川」という地名を一般的な地名照合システムで変換すると，地名辞書に「由良川」は含まれていないため，「京都府」に対応する領域または京都府内の任意の点（ほとんどの地名照合手法では便宜上府庁舎の位置を採用）に変換される．対象の魚自体が動くので採集地の誤差数メートル以内という高い精度は無意味だが，地図化されたデータから特定の水系付近にポイントされたものだけを検索するといった利用が考えられるため，どの水系で採集されたのか地図から判別できる必要がある．上の例では，変換した結果がたまたま桂川付近にポイントされてしまったとすると，由良川は日本海に注ぐ川，桂川（淀川水系）は瀬戸内海に注ぐ川であり，大きな違いがある．そこで，河川名を含む地名辞書を作

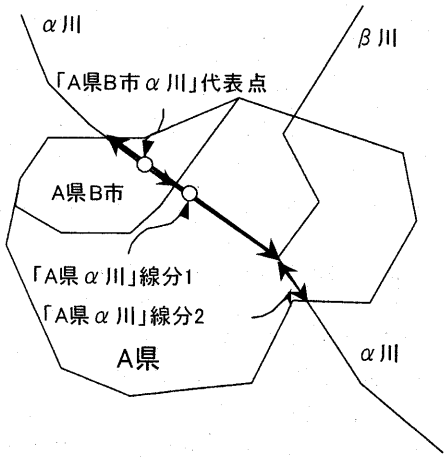

図 2 河川名称に対応する代表点

成することにより，「京都府由良川」のような曖昧な地名表記から京都府内の由良川付近の 1 点に変換することが可能な地名照合手法を開発する必要がある．

2.3 河川名を含む地名辞書の構築

一般に河川は幾何的にはネットワーク構造をもつ線の集合として扱われるが，地名照合では処理を簡単にするため，地名を領域や線分ではなく点に対応付ける．そこで，「住所＋河川名」という地名を変換できる辞書を作成するには，それぞれの地名表記に対応する点（代表点）を決定するルールを工夫する必要がある．住所と河川名を含む地名辞書を作成する手順を以下に示す（図 2）．

手順 1　合流・分岐による分割

同じ水系に属する河川でも，支流では呼び名が異なるということが多い（例：「由良川」は「淀川」の支流）．そこで，河川を合流および分岐点で分割し，それぞれの線分に対応する河川名を与える．

手順 2　自治体領域ポリゴンによる分割

河川は複数の自治体を通過するので，手順 1 で分割された線分を自治体ポリゴンで再び切断する．また，住所の詳細度に合わせるため，「県の領域ポリゴンで切断した線分」と，「市の領域ポリゴンで切断した線分」を作り，それぞれに「県・河川名」，「県市・河川名」を与える．

手順 3　最適な代表点の選択

合流しても名称が変わらない場合，同じ地名を持つ線分が複数存在する可能性がある．そこで，同じ地名を持つ線分のうち，最長となる線分を選択し，

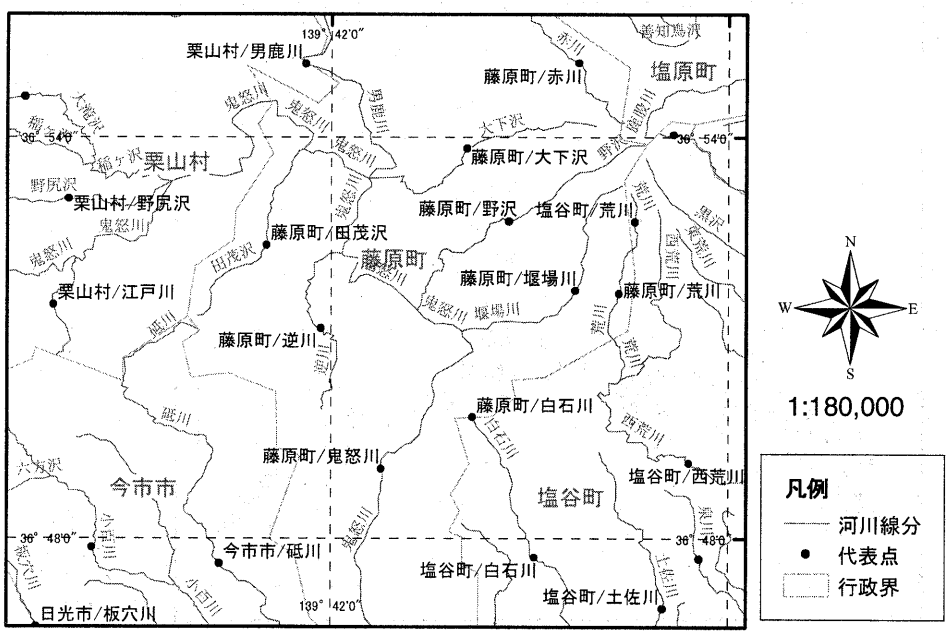

図 3 鬼怒川周辺の河川線分と代表点

その中心の座標値をその線分がもつ地名に対応する代表点とする．図2では「A県α川」に対応する線分が2つ存在するが，線分1が線分2より長いため，線分1の中心を代表点として採用する．

上記の手順によって河川名を含む地名辞書を構築することができる．図3に，藤原町鬼怒川付近の河川線分と市町村レベルに対応する代表点の分布を示す．図の左から下に流れる鬼怒川に注目すると，藤原町に入り大下沢，逆川，野沢，堰場川の順に合流する．中央付近の堰場川との合流点より下流部分の線分が最長になるため，この線分の中央点の座標が「藤原町鬼怒川」という地名に対応する．全国の名称を有する河川に対して本手法を適用することで23,986個の代表点を得た．

2.4 地名照合

標本データベースに含まれる地名表記は，先の例のように，河川名の位置が先頭であったり最後尾であったり，河川名と住所の間の区切り文字がカンマであったりコロンであったりと，明確な規則がない．しかし，前節で作成した河川名を含む地名辞書を利用するためには，標本データベースに含まれる地名の表記を地名辞書に合わせる必要がある．作成した地名辞書では，次の規則に従う表記を正規形とした．

規則1　住所部分は広い領域（県名）から狭い領域の順に記述する

規則2　河川名は住所の後に記述する

規則3　各項目の間は続けて記述する（セパレータは置かない）

この規則にしたがって前述の地名を正規化すると，次のようになる．

「広島県豊田郡本郷町沼田川本郷橋」

「茨城県鹿島郡波崎町太田地先利根川本流」

「埼玉県北葛飾郡幸手町權現堂川」

「北海道紋別郡雄武町御西川中流」

地名の正規化を行う処理は次の手順で行う．

手順1　もとの地名表記 g をセパレータの可能性がある文字（カンマ，コロン，空白など）で分割し，部分地名集合 $G = \{g_i, (i = 1..n)\}$ を得る．

手順2　G のすべての順列 $P = \{p_j, (j = 1..n!)\}$ に対し，部分地名を結合した文字列 s_j を作成し，地名照合を行う．

手順3　照合に成功した文字列長が最長となる s_j を正規化された地名として選択する．

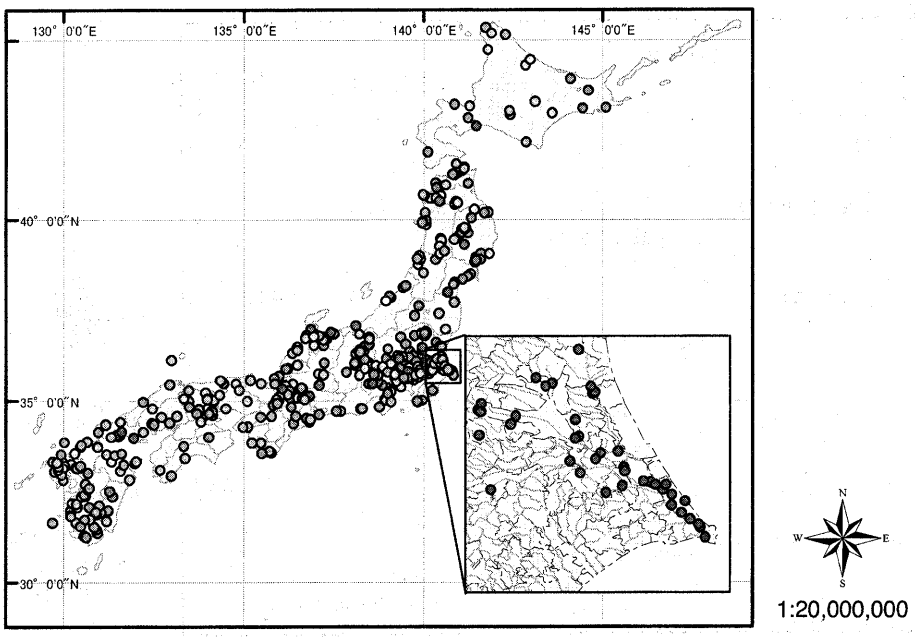

図 4 国立科学博物館魚類標本データの採集地分布

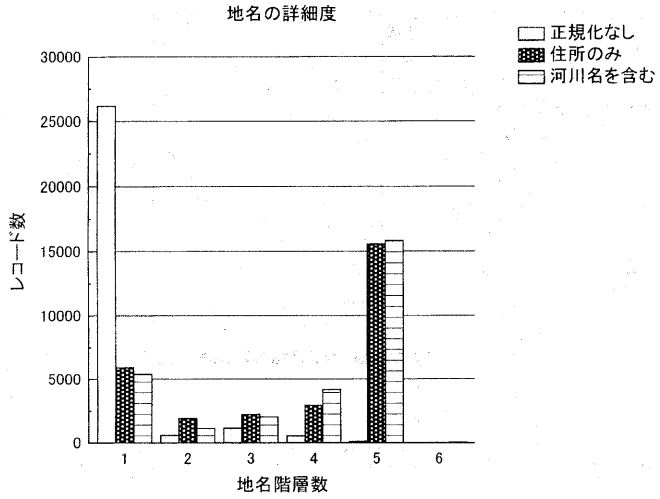

図 5 地名の正規化および地名辞書による地名照合結果の相違

以上の処理を行った後に地名照合を行うことで，標本データを最適な位置にポイントすることができる．地名照合には汎用的な地名照合サーバ SPAT[3]を利用した．対象データベースを地図化した結果を図 4 に示す．

2.5 地名照合結果の考察

地名の正規化を行わなかった場合，河川名を含む地名辞書を利用しなかった場合（住所のみ）と，河川名を含む地名辞書を利用した場合の地名照合結果の違いを図 5 に示す．地名階層数とは，地名照合で照合に成功したレベルを表す．例えば県名のみ照合できた場合，あるいは河川名のみ照合できた場合は地名階層数 1 となり，県名，市名，河川名が照合できた場合は地名階層数 3 となる．地名階層数が大きいほど正確な位置にポイントできていることになり，経験的に 3 以上であれば日本地図上に表示するには十分な精度（数 km～数 10km），4 以上であれば解析にも十分に正確な精度（数 km 以内）であると考えられる．

まず，グラフから地名の正規化が非常に大きな効果をもつことが分かる．地名の表記方法を統一すればこのような問題は起こらないが，既存のデータを活用するためには柔軟な照合手法が有効であることを示している．

次に，住所のみの場合に比べ河川名を含む地名辞書を利用した場合に階層数 1，2，3 のレコード数が減少し，4，5，6 のレコード数が増えていることが分かる．定量的には，実用上十分な精度であると考えられる 3 以上の地名階層数に変換できたものは，前者が 18,530 レコード，後者が 20,010 レコードで，それぞれ全体の 64.89%，70.08%であり，約 5%照合率が向上した．また，地名階層数が 3 から 4，4 から 5，5 から 6 に増加したデータが 1,480 レコードあり，これを含めると全体の 10.37%にあたるデータで照合精度が向上している．

また，グラフには表れていないが，河川名を利用することで階層数は増えないが位置精度が向上している場合もある．たとえば，「神奈川県川崎市多摩川」という地名は，河川名を含まない場合「神奈川県川崎市多摩区」にポイントされる．これは地名照合に利用した SPAT が持つ類似文字列を検索する機能によるもので，川崎市多摩区役所の座標値が返される．一方，河川名を含む地名辞書を利用した場合，同じ地名が「神奈川県川崎市多摩川」に変換され，多摩川上の一点の座標が返される．どちらの場合も知名階層数は 3 となるが，後者がより正確な位置にポイントされている．

以上の考察から，地名の正規化と河川名を含む地名辞書の構築が，魚類標本データベースに含まれる地名表記の地名照合に有効であるといえる．

3. 標本データベース地図ブラウザ

地名照合した結果である生物標本の分布図は，生物学・分類学での利用価値が高い．また，一般の利用者にとっても，馴染みのある魚の分布を閲覧するという楽しみ方がある．地理情報システムを利用すれば簡単に図化することができるが，一般に高価な上に使いこなすには時間がかかり，手軽に利用するには適していない．そこで，Web 上で標本データベースを閲覧するための地図ブラウズシステム「うお・まっぷ」の試作版を構築した（図 6）．

このシステムでは，サーバ上の魚類標本のデータベースから，和名・属名・種小名による検索が行える．検索結果は地図上に点として表示され(図 6 左)，それぞれの点をクリックすることで標本データ詳細

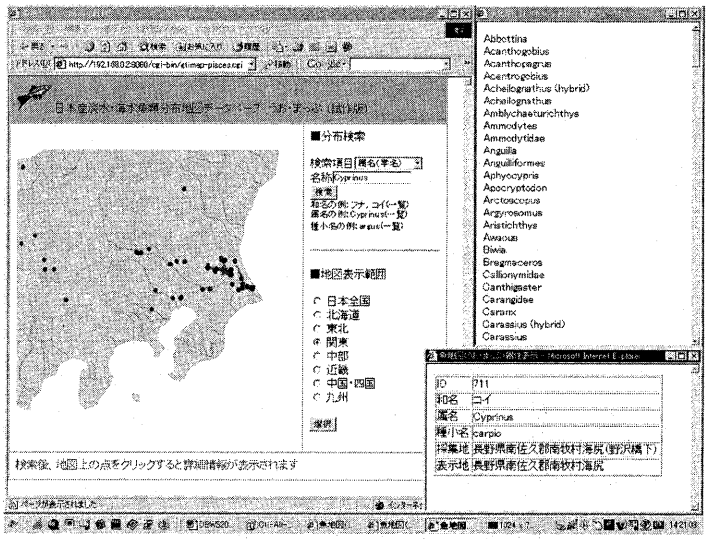

図 6 魚類標本データベース地図ブラウザ「うお・まっぷ」

を別ウィンドウで表示する（図6右下）．また，一般利用者の利用も想定し，データベースに登録されている和名，属名，種小名の一覧を表示する機能も持たせた（図6右上）．地図部分はSVG (Scalable Vector Graphics) [4]で描画しており，Adobe社が無料配布しているプラグインを利用して Web ブラウザ上で閲覧できる．また，簡単な拡大・縮小やスクロールも可能である．

4 まとめ

国立科学博物館所蔵の魚類標本データベースを，採集地地名をもとに地図化する一連の手順を説明した．規則性の低い地名表記を正規化する処理と河川名を含む地名辞書を作成することで，高い精度で地図上にポイントすることができた．他の分類群の標本データベースを構築する場合にはそれぞれに適した地名辞書を構築する必要があるが，本稿で示した地名から代表点を作成する手法を活用することが可能だろう．

また，地図化したデータを閲覧するためのブラウザを Web 上で構築した．このシステムはまだプロトタイプレベルだが，閲覧に必要な機能は満たしている．ユーザインタフェースを中心に改良を加え，近日中に一般公開を行う予定である．

現在の地名照合システムは日本語（漢字）で表記された地名しか扱えないため，海外の標本データや国内のローマ字で表記されている標本データを扱うことができない．海外の住所体系やアルファベット（ローマ字を含む）で表記された地名にも対応できるよう拡張することが今後の課題である．

謝辞 本研究は，環境省地球環境研究総合推進費の助成を受けて行ったものである．地名辞書の構築の際には，河川データは国土地理院国土情報整備室提供の国土数値情報・水文から，行政界データは同・国土骨格を利用した[5]．また地図データの処理にあたり，東京大学空間情報科学研究センター高橋昭子研究支援推進員に多くの助言をいただいた．

参考文献

[1] 岩槻邦男，多様性からみた生物学，裳華房，2002
[2] 野村総合研究所，GBIF 関連調査 調査報告書，2001，http://bio.tokyo.jst.go.jp/biores/siryo/gbif.pdf
[3] 相良 毅，有川 正俊，坂内 正夫，分散位置参照システム，情報処理学会論文誌，Vol.42，No.12，2928-2940，2001
[4] SVG: Scalable Vector Graphics，W3C，http://www.w3.org/Graphics/SVG/Overview.htm8
[5] 国土数値情報ダウンロードサービス，国土交通省国土計画局総務課国土情報整備室，http://nlftp.mlit.go.jp/ksj/