Ｆ－６　アジアオセアニア地域における生物多様性の減少解決のための世界分類学イニシアティブに関する研究

（２）GTI地域プログラム実施における生物多様性情報共有化と利用に関する研究

③　生物多様性情報の持続的構築に関する研究

東京大学

生産技術研究所　戦略情報融合国際研究センター　相良毅

平成１４～１６年度合計予算額　７，１３５千円

（うち、平成１６年度予算額　※　０千円）

「上記の予算額には、間接経費　０千円を含む」

［要旨］生物多様性情報の構築は、これまで主に標本データの収集、同定、管理に多くの人的・金銭的資源が割かれ、蓄積された標本情報の電子化・データベース化にはあまり積極的に取り組まれてこなかった。しかし近年、生態系への関心の高まりやインターネットの普及により、生物学の専門家ではない一般の利用者への生物多様性情報の公開が国際的に進められており、電子化・データベース化を持続的に行う必要がある。このような背景のもと、本研究では(1)電子化・データベース化による活用方法の提示、(2)電子化・データベース化における技術的な問題の軽減の２つを情報工学の手法を用いることで実現することを目的とした。(1)は生物多様性情報の構築を進める原動力を増し、(2)は生物多様性情報の構築にかかる障害を低くすることによりコストを軽減する。具体的には、標本ラベルに記載されている採集地の文字列情報を自動的に経緯度に変換する地名照合技術を用い、(1)に対しては標本データベースから生物種の分布図を自動的に作成し可視化するシステムを開発し、(2)に対しては標本ラベルのデータベース入力時に対話的な地名表記の正規化やエラー訂正を行うシステムを提案した。

［キーワード］生物種分布図、地名照合、インターネット、地図配信、入力支援

１．はじめに

生物多様性の理解と保全のためには、まず生物多様性情報を把握し、理解しなければならない。しかし生物多様性情報は地域的・組織的に分散して管理されているため、目的の情報を閲覧するだけのために多くの手続きと時間が必要となるのが現状である。この問題を解決するため、情報の電子化とインターネットを利用した統合化が国際的にも進められているが、そのための金銭的あるいは人的コストは決して無視できるほど小さいものではないことや、生物学・分類学の分野にはデータベース化やインターネットへの公開のための情報技術を持った研究者が少ないことが電子化を進める上での障壁となっている。特にアジアオセアニア地域は開発途上の国が多く、相対的に貧しい上に情報技術を持った技術者も少ないため、技術的支援なしには生物多様性情報の電子化を進めることが困難であり、この障壁を軽減することが、持続的な生物多様性情報の活用にとってきわめて重要である。

電子化の課題は、大きく２つに分けられる。１つは、生物多様性情報は大学等の研究機関や博

物館には存在しているものの、その記述様式が異なっているため、互換性がないことである。この問題については共通フォーマットの策定を中心に積極的な活動が進んでいる。もう１つの課題は、電子化にかかる金銭的あるいは人的コストを研究者個人または研究機関が負担する上で、明確なメリットがないことである。この問題は前者に比べて十分に認識されていないが、情報技術の手法を用いて電子化にかかる労力を極力軽減するとともに、電子化によるメリットをわかりやすい形で提供することが、持続的な生物多様性情報の整備に不可欠である。

２．研究目的

本研究では、アジアオセアニア地域における生物多様性情報の電子化にともなう問題を明確化し、情報技術にもとづいた新たな手法またはシステムを用いて、これらの問題を解決することで、持続的な生物多少性情報の活用を推進することを目的とする。初年度である平成14年度は、まず関連分野の研究者から問題を聞き取り調査することからはじめ、情報を整備する側にとっても生物多様性情報の電子化が有効に活用できるシステムの開発を目指した。

２年目である平成15年度は、初年度の調査の結果に基づき、生物多様性情報を積極的に活用するため、生物種の分布図を作成し、その時空間的な変化を可視化するシステムの構築を最終的な目標とした。そのためには、生物標本などの情報源に含まれる場所の情報が必要となるが、一般に場所の記載は地名などによる表記が用いられており、可視化する際には経緯度に変換する地名照合手法を確立しなければならない。地名照合手法については初年度中に開発を行い、日本語で記載された地名であれば語の順番が変わっていても経緯度に変換できるものを実装した。しかし英語（ローマ字）表記の地名が変換できないことや、そもそも生物多様性情報で用いられる地名が自然地名を中心としており、地名辞書の整備が不十分なことが課題として残っていた。そこで英語表記にも対応できる地名照合手法と、利用者による地名・経緯度情報の登録システムの開発を目的とした。

３．研究方法

（１） 生物標本採集地記載からの地名照合手法の開発

生物多様性情報の一つである生物標本は、比較的データベース化が進められている。しかし、GTI ワークショップなどでのヒアリング調査によれば1)、生物標本を保有している研究機関や博物館などでは、データベース構築に必要な技術を持つ情報処理技術者の不足や、生物標本ラベルの記載をコンピュータに入力するための人件費が十分ではないことなどが原因で、電子化が進まない場合がある。特に問題となるのは、ラベルの表記が記載者によって非常にばらつきがあり、かつ、手書きで読解が困難なケースがあるなど、基本的には単純作業でありながら、ある程度の専門教育を受けた人間でなければ入力作業も行えないことである。この問題を技術的に解決するのは非常に難しいが、専門家が多少の苦労を負担してでも電子化を行うことによって十分なメリットが得られれば、作業が進む可能性がある。

標本の電子化作業それ自体は、現状では学術的な業績として認められないため、電子化をおこなうことによって専門家が得られるメリットとしては、標本管理の手間を軽減することと、電子化されたデータを用いて新たな知見を得ることが考えられる。特に後者は情報処理技術の活用により、データのクリーニングや高度化といった可能性が考えられ、本研究の目的に合致する。そ

こで標本ラベルに記載されている採集地の情報に着目し、この情報から対応する経緯度情報を自動的に算出することにより、これまで作成が困難であった生物種分布図を自動的に作成する手法の開発を行った。

（2）インターネットによる標本分布図の公開システムの開発

生物多様性情報は生物学・分類学の研究者のみが利用するものではなく、最終的には一般市民が関心を持ち、日常生活の中で情報を活用していかなければ、生態系の保全は実現できない。そのためには生物多様性情報を一般市民にも理解しやすく、興味を持ちやすい形態で、かつ手軽にアクセスできるように公開しなければならない。一般市民の理解を得ることは、同時に生物多様性情報の取得や流通も促進し、正のフィードバックを生じる可能性がある。そのため、（1）で開発した生物種分布図を、インターネット上で手軽に公開・閲覧するシステムを開発した。

生物種分布図は、きわめて直感的に理解しやすく、身近に生息する生き物の調査や学習にも活用できるため、単に標本のラベル記載情報を一覧表として表示するよりも受け入れられやすい。また、標高など地理的な情報を重ね合わせることで、種の分布と生息環境をみるといった利用方法も可能である。

（3）言語によらない地名照合手法の開発

英語での地名表記には、ヘボン式ローマ字と日本式ローマ字による表記の違いや、記載者によって県がprefectureまたはprovinceと表現されるといった翻訳上の違い、スペルミスなど多くのバリエーションがあり、単純な文字列の比較では十分な変換精度が得られない。そこで、インターネット上のサーチエンジンをはじめとする全文検索手法で広く利用されているN-gram手法を拡張し、言語や文字に依存しない地名照合手法を開発した。本手法は確率的な手法であることから、多少のスペルミスや表現の違い、異なる文字が含まれていても、もっとも近似した候補を選ぶことができるため、結果的に正解を得られる可能性が高い。

（4）地名・経緯度対応情報入力ツールの開発

一般に地名は行政のために用いられる住所と、自然地形に与えられた自然地名に分けられる。住所は曖昧さが少なくデータの整備も進んでいるため、地名辞典(gazetteer)も整備しやすい。一方、自然地名については大きな地形に対する地名は電子化されているものの、狭い地域で用いられている地名は無数にあり、そのすべてをトップダウンに整備するのは不可能である。特に生物多様性情報については観測点がある程度決まっていることも多く、観測点に用いられる地名を登録できるツールを開発することで地名照合の精度を格段に向上させることができると考えられる。そこで地図上をクリックし、対応する地名を入力することで、その地点の経緯度を登録できるサーバシステムのプロトタイプを構築した。

4．結果・考察

（2） 地名照合手法

生物標本ラベルに記載されている地名は、一般的な郵便業務や役所業務で利用される住所を利用したものだけではなく、山や湖といった自然地名を含む自由な形式で記述されている。今年度は基礎的な調査のため、本プロジェクト参加機関である国立科学博物館所有の魚類標本データに含まれる採集地の情報を対象として、高い精度で対応する経緯度を求める手法を開発した。

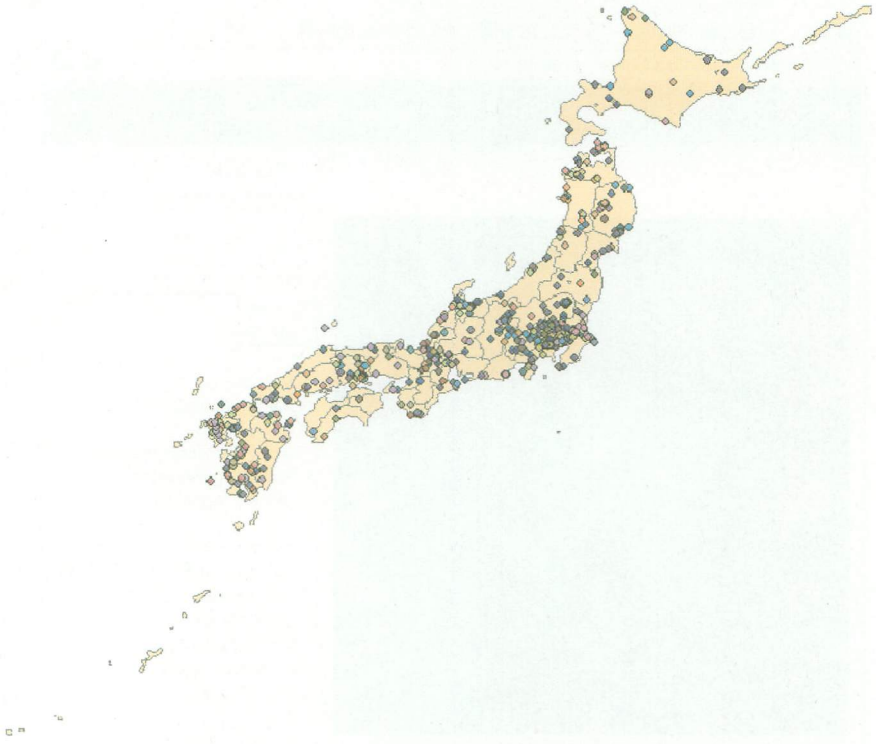

図1 魚類標本の種別分布

対象としたデータに含まれる採集地の表記では、市町村名や町字などの行政地名と、河川や湖沼といった自然地名を組み合わせたものが大半を占めているが、それぞれが出現する順序や区切り文字などは統一されていない。そこで、まずは自然言語処理で広く用いられる単語切り出しの手法を利用し、地名表記を単語に分解する。次に、得られた単語列を任意の順で並べ替え、地名辞書に含まれる地名と最長一致となる候補を選択することで、もっとも精度の高い経緯度を得ることができる。ただしこの手法では、地名辞書に含まれるレコード数が結果の位置精度に大きな影響を与えるため、ベクトル地図から詳細な地名辞書を作成する手法もあわせて開発した。

その結果、採集地表記のうち約70%から実用的な精度（数km精度）の場所情報を取得することに成功し、図１に示すような生物種分布図を自動的に作成することができた。このような分布図の作成は、これまでは個々の標本ラベルの採集地表記を人間が読み、地図の上に手作業でプロットするという膨大な作業が必要であったため、手軽に行うことができなかった。本手法を用いることで分布図の作成が容易になるため、標本データの電子化の一つの動機となるだろう。

（２）インターネット分布図配信システム

標本の分布図をインターネット上で閲覧可能なシステムを開発した。分布図を配信する場合、あらかじめサーバシステム側でいくつかの代表的な画像を作成しておき、ユーザがその中から選択してWebブラウザ中に表示させるタイプのものや、ユーザ側で学名や通称名を用いて検索を行い、動的に画像を生成してWebブラウザ中に表示させるものが代表的である。しかし、これらのシステムではユーザ側で拡大・縮小といった処理を行うたびに画像の再検索・再受信を行わなければならず、操作性が低いという問題がある。また、インターネットGISと呼ばれる商用システムをサー

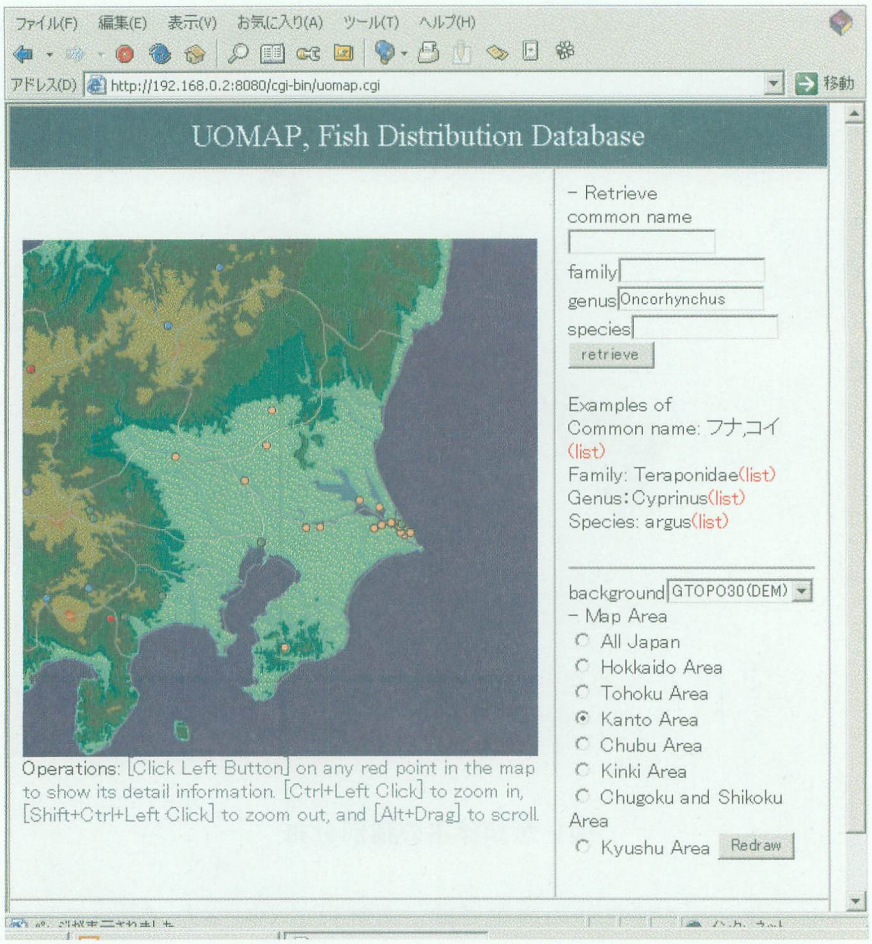

図2 インターネット分布図配信システム

バ側に利用した場合、サーバ構築に多額の維持コストが必要となり、継続的にサービスを提供するのが難しいという問題がある。

これらの問題を解決するため、今回開発したシステムでは、地図画像の部分にSVG（Scalable Vector Graphics）[2]技術を利用した。SVGはWebの標準を定める国際機関であるW3Cで勧告が出されており、事実上の世界標準として広く利用されているベクトル画像のフォーマットである。SVGを利用することにより、ユーザ側で拡大・縮小・スクロールといった処理を自由に行なうことが可能になる。また、サーバシステムは一般的なWebサーバにデータベースを組み合わせるだけで済み、維持コストがほとんどかからないというメリットもある。また、印刷時にラスタ画像特有のジャギーが発生しないことや、画像をダウンロードして商用のグラフィックソフトで加工し、論文など印刷物に貼り付けて利用できるという、再利用に際しての利点も大きい。その一方で、ラスタ画像に比べて画像の生成やデータ転送に時間がかかることが問題になる可能性が考えられたが、サーバ側で線分の間引き処理を行うことで、利用上全く問題のない速度が実現できた。また、図２のように標高データなどのラスタ画像を背景に重ねることも可能であり、幅広い応用が考えられる。

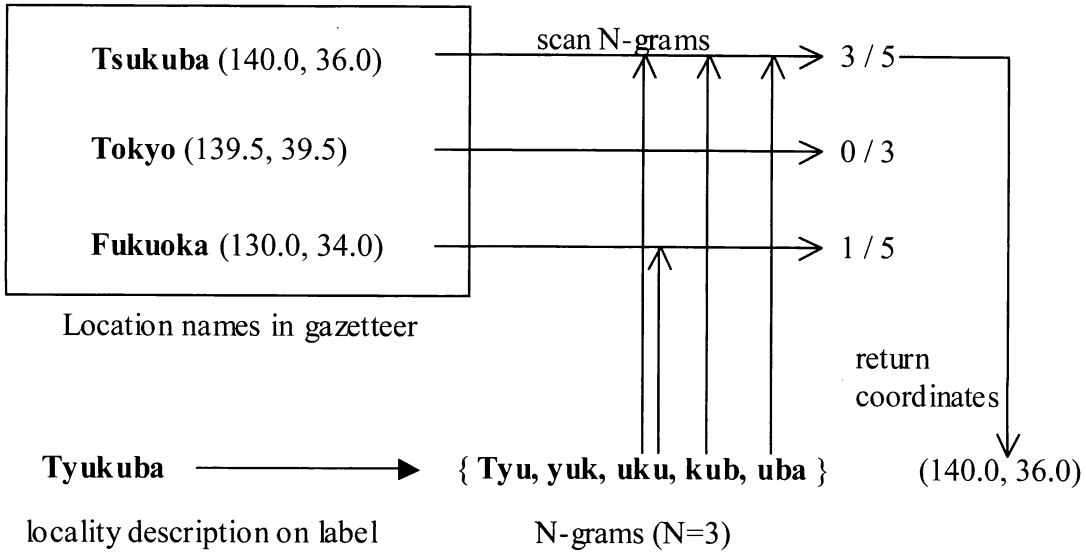

図3. N-gram を用いた地名照合手法の概要

（３）言語によらない地名照合手法

生物標本ラベルに記載されている地名は、一般的な郵便業務や役所業務で利用される住所を利用したものだけではなく、山や湖といった自然地名を含む自由な形式で記述されている。初年度開発した手法は語の順序が変化しても対応できるもので、日本語で記載された地名であれば（その地名が地名辞典に登録されている限り）経緯度に変換できた。しかし英語で記載された地名は、ヘボン式と日本式ローマ字のどちらが採用されているかによりスペルが変わってしまうため、対応できなかった。たとえば「つくば市」は英語表記の場合「Tsukuba City」がもっとも一般的だが、日本式表記を採用すれば「Tukuba City」に、「つくば市」そのものが固有名詞であると考えれば「Tsukuba-shi City」あるいは「Tukuba-si」と表記されることもある。これは日本固有の問題ではなく、英語以外の言語を母国語とするすべての国にとって共通の問題であり、この問題を解決することは国際的にも大きな貢献である。

開発した手法は、N-gramを用いた文字列間の「類似度」を用いるものである。図３は、地名辞典にTsukuba, Tokyo, Fukuokaが登録されているとき、タイプミスを含む“Tyukuba”が検索された場合を想定したものである。従来の地名照合手法では単語同士の比較を行うため、タイプミスが存在すると完全に不一致と判断してしまう。これはタイプミスだけではなく上述したローマ字表記の違いでも同じである。提案手法では、文字列をN文字ずつの文字組ととらえ、何組の文字組が一致したかを数えることによって比較を行う。そのため、１文字違うと最大でN組が不一致となるが、それ以外の組は影響を受けない。図３ではN=3の例であり、3組が一致する“Tsukuba”が検索語“Tyukuba”にもっとも近い結果として返されていることがわかる。

一方で、あらかじめすべての地名から文字組のインデックスを作成しなければならないことや、すべての地名と比較しなければ最適解が求められないため処理が遅いという短所もある。処理速度については、(a) 頻繁に出現する語（都道府県名など）はグループ化する、(b) 枝刈り手法による比較対照の絞り込みを行うという２つの技術的工夫により、十分に実用的な処理速度を実現することができた。

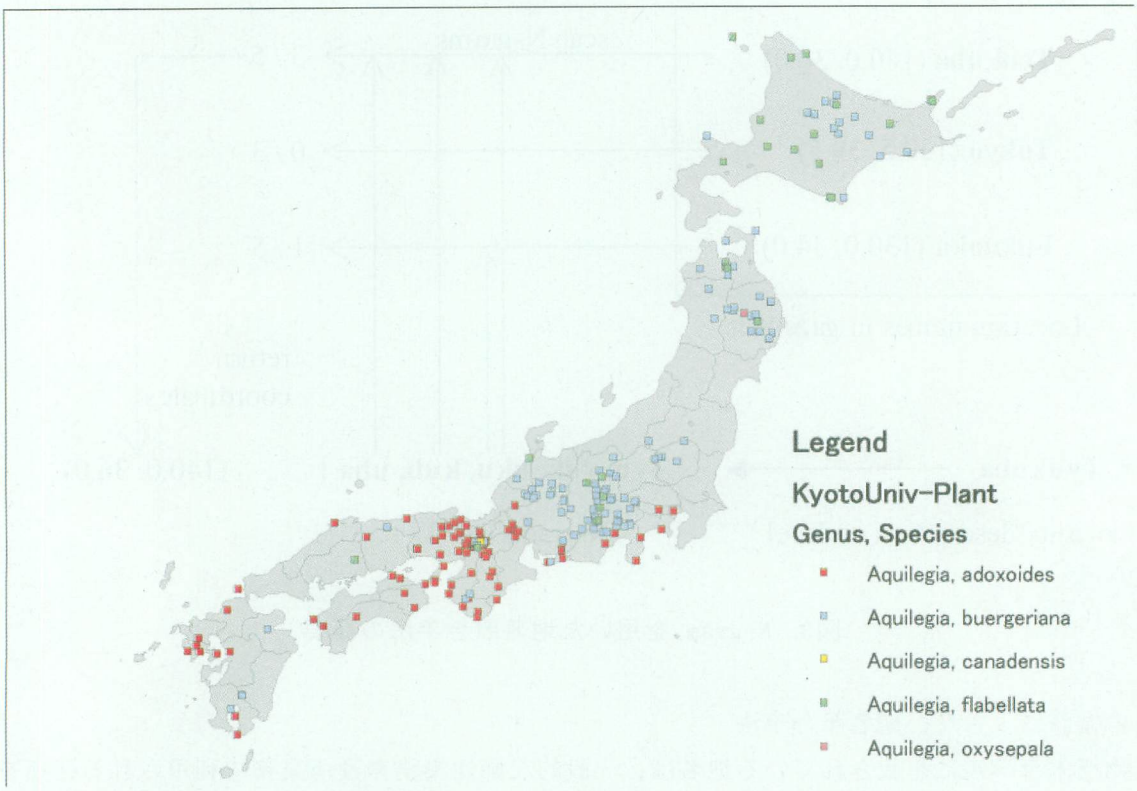

図4. 提案手法による地名照合の結果（英語地名に対応）

図４は、本研究プロジェクトに参加している京都大学戸部教授のグループによる植物標本のデータベースを、提案手法により地名照合を行い、地図化したものである。このデータベースは地名の記載が英語であったため、初年度の手法では十分な変換精度が得られなかったが、今回の手法では約70%と比較的良好な精度で変換を行うことができた。なお、このデータベースに対して上記(a), (b)の工夫を行わなかった場合と行った場合の処理に要した時間を比較したところ、約120分の1に処理時間が短縮された。

提案手法はアルファベットにも依存しないため、あらゆる言語に対応できる点や、語の順番にも影響されない点が長所である。しかし、確率的な手法であることから、変換結果は「何%が一致した」という形でしか得られないため、最終的に人間が上位候補群より正解を選択して判断する必要があり、一括処理には向かないという欠点もある。そこで、対話的な変換ツールも試作して性能を測定したところ、処理速度の向上により、１秒以内で応答するシステムを実現できた。

（４）地名・経緯度対応情報入力ツールの開発

地名照合手法を実際に活用するには、地名辞典の整備が不可欠である。現在、国際的にもデジタルライブラリプロジェクトの一環としていくつかの地名辞典の整備が行われている。しかし、これらの活動は世界中の地名を網羅的に収集するため、より地域性の高い小規模な地名については十分にカバーできないという制約がある。特に地名辞典の整備を主体的に行っているグループは地理学を専門とするため、生物多様性情報の地名記載に頻繁に現れる地名とは焦点が異なっており、地域や専門によって独自の地名辞典を整備する必要がある。

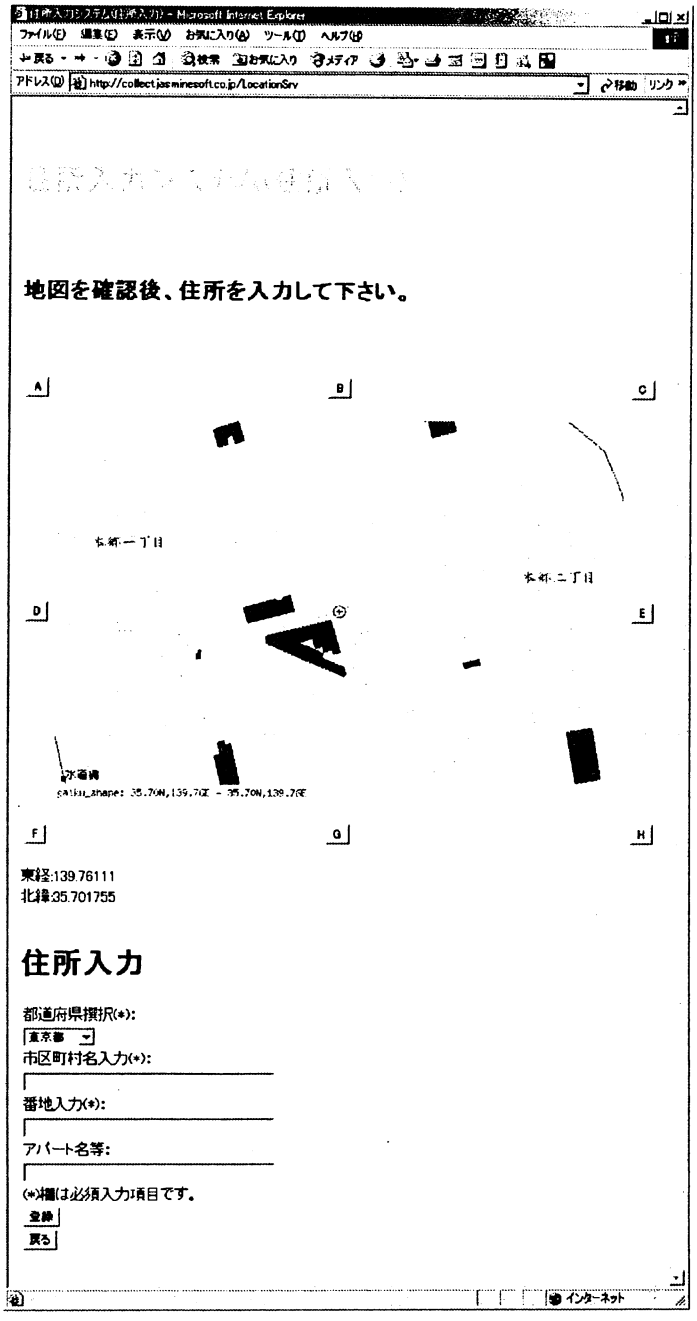

図5. 地名・経緯度対応情報入力ツール

この問題を解決するためには、利用者が積極的に地名と経緯度の対応を入力できるツールが有効である。その際、経緯度を値で入力できる利用者は少ないため、地図上で場所を与えるインタフェースが求められる。図５は試作した入力ツールである。このツールで一度登録した地名は地名照合手法で利用できるため、定点観測を行うような地名を登録することにより、地名照合の精度を向上させることができる。

５．本研究により得られた成果

これまで生物多様性情報の整備は、情報を保有する側の善意や義務感に頼って進められてきた

側面があり、これが持続的な整備の上で大きな問題となっている。本研究では、情報を保有する側にとっても情報整備が直接的なメリットとなる例を示すため、(1)電子化された標本データから分布図を自動作成する手法と、(2)インターネット上で簡単に公開するシステムを開発した。これらの手法・システムは、国内外の標本データの電子化の動機の一つとなり、情報の公開が進展する可能性があるという点で重要である。また、(3)英語（ローマ字）で表記された地名や海外の地名にも対応する地名照合手法を開発し、実際に英語で記載された生物標本データベースを地図化してその有効性を確認した。最後に、(4)地名辞典に記載されていない地名を利用者が独自に登録できる地図ベースの入力ツールを開発した。

ただし、これらの手法を生物多様性情報の持続的な構築に対して有効に機能させるためには、登録された地名を利用者間で共有できる仕組みを備えた、持続的な地名照合サービスの提供が不可欠であることも明らかになった。

６．引用文献

1) Shimura, J., ed. 2003. Global Taxonomy Initiative in Asia: Report and Proceedings of 1st GTI Regional Workshop in Asia, Putrajaya, Malaysia, September 2002. v+314 pp. National Institute for Environmental Studies, Japan.

2) SVG Working Group: Scalable Vector Graphics, http://www.w3.org/Graphics/SVG/Overview.htm8

７．国際共同研究等の状況

（１）研究計画名： Human Resource Development Program: Asia-Pacific Telecommunity

協力案件名：Communication Between Online Heterogeneous Repositories : An Application of Simple Object Access Protocol (SOAP) for Rapid Knowledge Discovery

カウンターパート氏名・所属・国名：Amir F. Merican (Institute of Biological Sciences, University of Malaya, Malaysia)

参加・連携状況：マレーシアの微生物データベースの分散構築に向け、データベースの設計および通信方式に関する検討を共同で行い、プロトタイプシステムを開発した。

国際的位置づけ：本プログラムはアジア・太平洋サミットによって提言された、この地域での情報通信技術の発展と人材育成を目的としており、本研究はその一環として、微生物データという実データを利用し、かつ実際に利用可能なデータベース・データ流通システムを開発したという点で重要である。

（２）研究計画名： The Ocean Biogeographic Information System (OBIS)

協力案件名：海洋生物多様性情報の整備について

カウンターパート氏名・所属・国名：Mark Costello (Huntsman Marine Science Centre, Canada)

参加・連携状況： OBISの次年度研究計画作成会議に参加し、生物多様性情報の電子化について意見を交換した。

国際的位置づけ： OBIS は世界最大の海洋生物多様性情報とりまとめ機関であり、海洋国である日本にとっても非常に重要な活動の１つである。特に、生物多様性情報を地理情報と関連づけて整備することを明確にうたっており、本研究における地名照合手法や地図化の技術とは関係が深い。

８．研究成果の発表状況

（１）誌上発表（学術誌・書籍）

<学術誌（査読あり）>

なし

<学術誌（査読なし）>

①．相良　毅、松浦啓一、佐藤　聡、志村純子：日本データベース学会Letters、1, 1, 39-42(2002)「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」

②．Sagara, T., Matsuura, K., Shimura, J., Research Report from the National Institute for Environmental Studies, 175, 281-286(2003), "A Web-based Biodiversity GIS Using a Robust Geocoding Algorithm."

③．Sagara, T., Building Capacity in Biodiversity Information Sharing 2003, 162-166 (2003), "An Efficient Address Matching Algorithm for Locality Descriptions on Specimen Labels."

<書籍・報告書類等>

なし

（２）口頭発表

<オーラル>

①．松浦啓一、林　洋平、瀬能　宏、相良　毅、志村純子：2002年度日本魚類学会年会（2002）、「統合型魚類データベースを目指して」

②．Sagara, T., Matsuura, K., Shimura, J. GIScience2002, Boulder/USA (2002), "A Web-based Biodiversity GIS Using a Robust Geocoding Algorithm"

③．Sagara, T., Invited Talk for 1st International Workshop on Agrobiodiversity, Pilot Conservation Project and Taxonomy Gap, Hanoi/Vietnum (2003), "A prototype specimen database using GIS"

④．Sagara, T., 1st GBIF Science Symposium, Copenhagen/Denmark(2003), "Cleaning and adding value to inaccurate geographical descriptions on specimen labels"

⑤．Satara, T. Joint International Forum on Biodiversity Information, Tsukuba/Japan(2003), "An Efficient Address Matching Algorithm for Locality Descriptions on Specimen Labels."

<ポスター>

①．Sagara, T., Matsuura, K., Shimura, J. 1st GTI Regional Workshop in Asia, Putrajaya/Malaysia (2002), "A Web-based Biodiversity GIS Using a Robust Geocoding Algorithm"

（３）出願特許

なし

（４）受賞等

なし

（５）一般への公表・報道等

なし

９．成果の政策的な寄与・貢献について

①　本研究の成果であるインターネットによる生物種分布図の配信システムは、下記URLで利用可能である。

http://spat.csis.u-tokyo.ac.jp/~sagara/cgi-bin/uomap.cgi

本システム上に多くの生物種データが蓄積されれば、環境保全や都市開発の際の参考情報として大いに利用価値がある。