データベースシステム 119−37
（１９９９．７．２２）

XMLを基本としたテキスト空間情報ベース

相良 毅† 有川 正俊† 高橋 昭子†

地理情報を利用するハードウェア，ソフトウェアの普及が進んでいるが，利用できるデータが少ないという問題がある．本論文では，個人で情報発信を行うような一般ユーザでも簡単に記述できる，空間タグ<spa>表現を提案する．地名しか含まない最も単純な<spa>表現でも，アドレスマッチング手法を利用して自動的に緯度経度情報を埋め込むことが可能で，十分実用的であることを示す．また，<spa>表現を持たない既存のデータに，自然言語処理を利用して自動的に<spa>表現を埋め込んで利用する『芭蕉』システムを説明し，応用例を示す．

Text Spatial Information Base Using XML Description

SAGARA Takeshi† ARIKAWA Masatoshi† TAKAHASHI Akiko†

It is becoming much popular to use hardware/software for geographic data, but useful geographic data are not sufficient. This paper proposes a spatial tag ,<spa> expression, which enables ordinary users to publish their personal spatial information. We discuss a framework that "Address-Geo-Coding" converts non-geo-coded description, such as "Shibuya", into geo-coded description, such as "(139.69641,35.66655)", in <spa> statements which are parts of statements and are defined by <spa> expression. Our prototype system "BASHO" is also presented to clarify how effective <spa> expression works using two real application samples.

1. はじめに

WWWやインターネットニュース，電子メールなどでは，ラーメン屋リストやスキー場案内など場所を含む情報を会社・個人レベルで発信している例は多い．住所などを特に意識せず記載している情報は更に多いが，これらの情報はHTMLやplain textなどで記述されており，有効に活用されていない．特に場所の記述は「～駅前」や「～町」といった自然言語で表現されているため，緯度経度のような構造化されたX-Y情報しか扱えないGISソフトでは利用できない．

一方，このような場所の情報を利用する，カーナビゲーションシステムやPC用電子地図ソフト，GPS装置などが手軽に利用できる程度の価格になり急速に普及が進んでいる．これらのシステムは，ベースとなる地図データや各種サービス情報を購入したり受信して利用するため，地理情報に対するニーズがますます高まっている．しかし，データの入力や編集，加工が可能な高性能な地理情報システム（以下GIS）ソフトは，投影法やデータの種類，操作などの知識を必要とするため，一般ユーザが個人的に情報発信するためには利用できない．

本論文では，個人発信情報など，そのままではカーナビやGISソフトでは利用できない情報を，地理情報として利用できる形で提供するために，単純なXML表現を利用した空間タグを埋め込むことを提案し，その有用性を評価する．また，空間タグを自動的に埋め込むシステムを作成し(図 1)，応用例として地理情報サーチエンジンを実装した．

†東京大学空間情報科学研究センター
Center for Spatial Information Science at the University of Tokyo

2. 地理情報の特徴

2.1 地理情報の現状

地理情報とは，狭義には，場所と形状を表す位置情報と，その位置に関連付けられたその他の属性情報を管理しやすい形に構造化したものである．多くのGISソフトでは，位置情報を2Dないし2.5Dの空間データ構造で管理し，その他の属性情報をリレーショナルデータベースで管理する．両者の間はシステム内で定義されたIDによって関連付けられる．現状では形状の表現やデータフォーマットがシステムに依存しているため，データベース間の相互操作性が乏しく，ISO TC211 および Open GIS Consortium（OGC）で国際的な標準化作業が行われている[1]．特に，OGCではXMLベースの地理情報記述を議論している．このような厳密に定義された地理情報の記述は，位置や情報の精度が高く利用価値も高いことが期待されるが，必要な入力項目も多くデータ作成が困難である．

さて，地理情報をより広義に「位置に関連する情報」として捉えると，場所を説明する簡単な文や略図を含む情報も地理情報と考えることができる．このような情報の例としては，各種店舗の広告，テレビや文字放送で流れる気象情報・交通情報，新聞記事，雑誌のタウン情報など，非常に多く存在する．これらの広義の地理情報は，GISソフトで利用できるように構造化されていないため，そのままでは二次利用することができず，各メディア内に散在している[2]．

さらに，位置の表現は非常に多様で，かつ主観的・概念的なことも多く，定型化が難しいという問題がある．例えば「渋谷駅ハチ公口より徒歩5分」のような表現から，渋谷駅周辺の地図上に正確な位置を定めることはできないが，位置に関するかなり詳しい情報が含まれている．これを定型化・構造化されたデータとして管理しようとすると，誤差精度の表現など詳細な定義が必要になる．しかし一般にはそれほどの正確さを必要としない用途もあり，より簡単で手軽な表現で十分なことも多い．

そこで，精度は低いが大量に存在する広義の地理情報を地理情報システムで利用可能にするための方法として，半構造化表現であるXML文法を利用した<spa>表現を提案する．

2.2 <spa>表現

<spa>表現は，個人的な情報発信者のような，GISの専門家ではない広義のデータ作成者でも簡単に利用できるように，きわめてシンプルな表現を許した

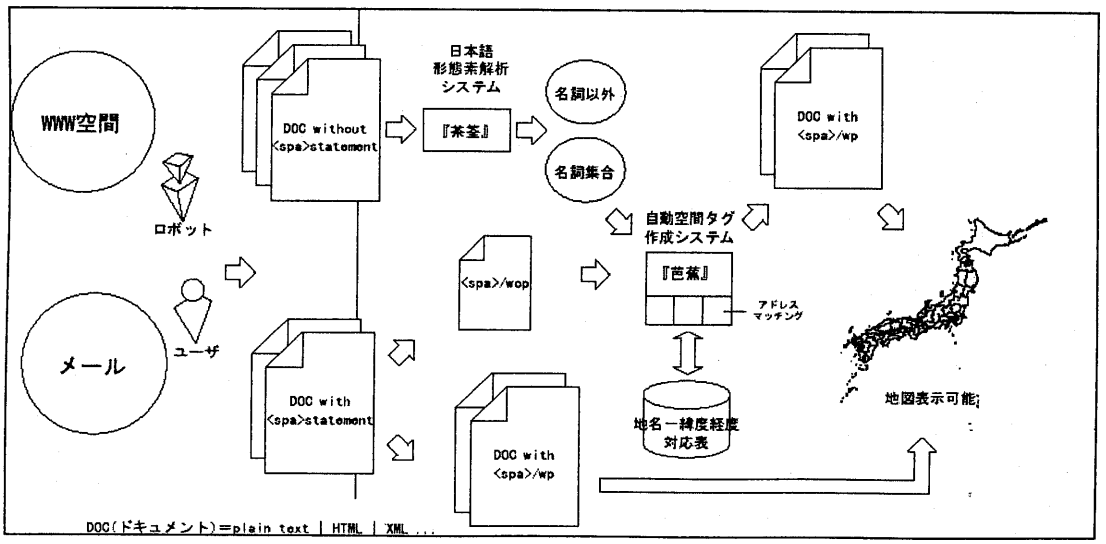

図 1　システム構成図

地理情報表現である.最もシンプルな<spa>文章は次のようになる.

<spa name="渋谷"/>

場所を表す表現は，その性質上，他の文書中に埋め込まれることが多いため,<spa>はXMLの文書要素(DOCTYPE)ではなく,他の文書要素中の一要素として利用されることを想定している．例えば，会議の案内文書で会場を説明する部分を<spa>及び</spa>で囲むことで，元の文章に影響を与えることなく場所を表すことができる．

次に，より詳細な表現が可能な<spa>表現の完全なDTDを図2に示す．次の例を用いて，各要素を説明する．

・・・次回の会議は
<spa name="東京大学空間情報科学研究センター会議室" type="point" p="139.83E 35.614N">
東京大学空間情報科学研究センター会議室
<geoword>東京大学</geoword>
<domain>東京都</domain>
<domain>目黒区</domain>
<domain>駒場</domain>
<author>相良 毅</author>
</spa>
で×月○日△時より・・・

<spa>内のname属性によって，これから説明する地理情報が「東京大学空間情報科学研究センター会議室」に関するものであり，type属性で形状が点(point)であることを表す．また，p属性で緯度経度が139.83E, 35.614Nであることを示している．p属性があれば地図上にポイントすることが可能であり，これ以降特に区別する必要がある場合，p属性を持たないものを「<spa>/wop」，持つものを「<spa>/wp」と表記する(“wp”と“wop”はそれぞれ“with-p”と“without-p”の略).

次のgeoword要素は，spa要素のp属性が実際に代表している情報の名称を表す.つまりこの例では，139.83E, 35.614Nという座標は「東京大学」の代表点であって，「空間情報科学研究センター」の正確な位置ではない．これに関しては後述するアドレスマッチングの項で詳しく説明する．

一連のdomain要素は，この地理情報が属している地理的な領域を表現する．同名の建物や町名などが存在するため，場所を特定するためのヒント的な情報を付加できる．

author要素はこのデータを作成・入力した著者，または自動生成したプログラム名を入力する．悪意を持ったユーザを排除したり，特定のプログラムの出力だけを利用するなどの用途が考えられる．

2-3 アドレスマッチングによる<spa>表現の変換

一般のユーザは緯度経度を入力できない可能性が高いが，アドレスマッチング手法を利用すれば，地名しか入力されていない<spa>/wop 文章を，自動的に緯度経度を含む<spa>/wp 文章に変換することができる．<spa>/wp文章は，地図上に表示したり分

```
<!ELEMENT spa (#PCDATA | geoword? | domain* | author? | longtude | latitude) >
<!ATTLIST spa name CDATA     #IMPLIED
              type (point | line | area) "point" #IMPLIED
              p    NMTOKENS  #IMPLIED >
<!ELEMENT geoword (#PCDATA) >
<!ELEMENT domain  (#PCDATA) >
<!ELEMENT author (#PCDATA) >
```

図 2 <spa>表現のDTD

布を統計的に解析するといった各種の空間処理・解析が可能になり，利用価値が高い．

アドレスマッチング手法は，一般に，地名と緯度経度の対応表をあらかじめ用意し，地名を検索キーにして緯度経度を検索する方法である．この手法を利用すると，<spa>/wop 文章を後処理によって自動的に<spa>/wp 文章に変換できる．地名以外にも電話番号，郵便番号などからも（対応表さえ利用できれば）緯度経度を追加することが可能である．

しかし，次のような技術上の問題がある．

・より詳しい地名を扱えるようにする程テーブルが大きくなり処理時間がかかる
・同名の地名があった場合どちらかを判定できない（キーが一意ではない）

そこでテーブルを都道府県／群・政令指定都市の単位で分割し，階層構造を構築してマッチングを進める．また，完全にキーに一致するレコードがなくても，マッチできた範囲で代替結果として利用する．例えば「東京大学空間情報科学研究センター」が存在しなくても，「東京大学」がマッチすればその点を代表点として採用する．

絞り込みのプロセスと不完全マッチの結果は，精度や内容を検証する上で重要な情報になるので，domain および geoword 要素として<spa>文章中に保存する．また，特定の domain だけを対象とした，より詳細なアドレスマッチングを行う対応表が存在すれば再度アドレスマッチングを行うことで精度を高めることができるが，その際にも domain 及び geoword が有用である．

1235	石狩市		141.3075	43.21442
1102	北区	札幌市	141.3477	43.13329
1407	仁木町	余市郡	140.70653	43.09655
1217	江別市		141.54482	43.10162
1422	栗沢町	空知郡	141.86435	43.14879
1103	東区	札幌市	141.40039	43.09665
1631	音更町	河東郡	143.20638	43.05827
1649	浦幌町	十勝郡	143.66988	42.95721
1403	泊村	古宇郡	140.524	43.1002
1109	手稲区	札幌市	141.22568	43.10012
1463	占冠村	勇払郡	142.51288	43.02142

図3 アドレスマッチング辞書

3. 自動<spa>文章作成システム

3.1 アドレスマッチング辞書

アドレスマッチングを行うためには地名・緯度経度対応表を管理する辞書が必要である．今回利用した辞書は，全国の都道府県・市町村ポリゴンデータ「全国市区町村界データ（株式会社パスコ）」を元に作成した．このデータでは，市町村の境界線がポリゴン（閉多角形）形式で記述されている．そこで，それぞれのポリゴンから代表点となる内部の点を計算し，緯度，経度，名称からなる一覧表を作成した．データ件数は 4118 件である．

また，それぞれの市町村には属する群・政令指定都市の情報が，群・政令指定都市には属する県の情報が記述されている．そこで，この階層構造を RDB のスキーマとして定義し，データを投入した．データの一部を図3に示す．

3.2 自然言語処理を利用した<spa>文章作成システム『芭蕉』

（原文）
東京，下北沢の魅力はなんといても新宿からは小田急線，渋谷からは井の頭線でわずか 10 分たらずで到着してしまうアクセスのよさ．

↓

（芭蕉による変換後）

東京，下北沢の魅力はなんといても新宿から<spa name="新宿" p="139.72015,35.70287"><geoword>新宿区</geoword><domain >東京都</domain><author>BASHO Ver.0.4</author></spa>は小田急線，渋谷から<spa name="渋谷" p="139.69641,35.66655"><geoword>渋谷区</geoword> ...

図4 『芭蕉』による自動<spa>文章作成例

既存のデータには<spa>情報が埋め込まれていないが，自然言語処理とアドレスマッチング手法を組み合わせることで，自然言語で記述された文章から位置情報を抽出することができる．

地名は厳密に言えばすべてが固有名詞であり，完全な固有名詞辞書を持つ自然言語処理システムがあれば完全な地名抽出が可能であるが，実際には日本中の地名をすべて登録した辞書は存在しない．しかし，地名は一般名詞と一般名詞，一般名詞と固有名詞を組み合わせた形になっていることが多い（例：下＋北沢＝下北沢）．そこで，形態素解析システムにより文章を品詞に分解し，１つ以上の名詞が連続して現れた場合には，その組み合わせをアドレスマッチング辞書で検索を行うという処理を行うと，形態素解析辞書に存在しない地名も抽出することができる．

今回開発したシステム『芭蕉』は，日本語形態素解析システム『茶筌』(奈良先端科学技術大学大学院大学)version2.0[3]の出力結果を利用して地名を抽出する．同時にアドレスマッチングを行い，緯度経度を含むspa情報を生成してオリジナルの文書に埋め込む処理を行う．

図４に『芭蕉』による自動<spa>文章作成の結果を示す．

4. 応用例

4.1 WWW 巡回ロボットによる地図作成システム

<spa>表現及び『芭蕉』の利用例として，WWW巡回ロボットを利用して既存のテキスト情報を収集し，『芭蕉』で<spa>表現を自動的に埋め込んだデータベースを作成した．『芭蕉』はアドレスマッチング手法により<spa>/wp情報を作成するので，地図上に表示したり，GISソフトで利用することができる．

一例として，各ページのURLと<spa>表現に含まれる緯度経度情報を利用して，クリッカブルマップを作成した（http://www.csis.u-tokyo.ac.jp-/spa/cgi-bin/onsen_map.cgi)．画面を図５に示す．

今回利用したページは「全国温泉案内『Ｏｈ！ＹＵ』」のトップページ（http://www.csn.co.jp/-ohyuhm.htm）からリンクされていて，かつ同ドメイン内に置かれている592ページである．そのうち288ページが各温泉の紹介となっており，残りは県別のインデックスページやホテルへのリンク情報などである．作成したクリッカブルマップでは，１ページ中に５件以上の地名が埋め込まれたページ（おそらくインデックスページと思われる）へのリンクを黄色，５件未満のページへのリンクを赤色のシンボルで表現している．

このシステムで，WWW空間に蓄積されている各種地理情報を収集し，高次利用することが可能になる．また，ニュース記事のページなどで定期的に動かすことにより，最新の情報を常に地図上にポイントする「地図版サーチエンジン」として利用することも可能である．

4.2 メールベース地図 BBS

もう一つの例として，位置情報を含むメールを『芭蕉』で処理し，地理情報として利用するシステムを開発した．当センター内に作成したメールアドレス宛に<spa>表現を含むメールを送ると，<spa>/wop 文章の場合アドレスマッチングを行い<spa>wp 文章に変換してから，自動的に地図上にポイントする．また，<spa>表現を含まないメールが送られると『芭蕉』により位置情報を埋め込んで地図上にポイントする．

将来的には災害や交通，気象などの情報を通報するシステムとして利用したり，生態系調査や水質調査など他分野の研究者が地理情報を収集するためのプラットフォームとしての応用を検討している．

5. おわりに

GIS ソフトの知識を持たない一般の利用者でも地理情報を容易に記述できる<spa>表現を提案した．この表現は，地理情報のもつ曖昧さや多様な表現も扱うことが可能で，XML の文法に準拠している．

次に，地名しか持たない<spa>表現に，アドレスマッチング手法を利用して自動的に緯度経度情報を追加するシステムを示した．都市部などより詳細な地図が利用できる地域や，大学内など構内地図を持つ場所では，さらに高精度なアドレスマッチングを行うことができる．

また，<spa>表現が埋め込まれていない既存のデータに対して，自然言語処理とアドレスマッチングを利用し自動的に<spa>表現を追加するシステム『芭蕉』を開発した．今回利用した『茶筌』の日本語辞書は 67420 件の地名を含んでおり（ただし海外の地名を含む），市町村名レベルの地名をほぼ網羅した非常に優れたものである．そのため『芭蕉』に渡される名詞もかなり精度が高いが，可能性のある名詞の組み合わせを一つ一つ問い合わせるため処理時間がかかる．<spa>表現があらかじめ埋め込まれていればこの処理を省略することができ，より高速な処理ができるが，<spa>表現の普及と並行して『芭蕉』の高速化も進めていきたい．

図 5 地理情報サーチエンジンの実行画面

今回のシステムではアドレスマッチング辞書として全国レベルの小縮尺データを利用したが，東京都市部などの大縮尺データを利用して，更に精度の高いアドレスマッチングを行っていく予定である．また，収集したデータを有効に利用するには，位置だけではなくキーワードによっても分類する仕組みも必要である．この仕組みも検討・開発していく．

謝辞

本研究を進めるにあたり，貴重な助言を頂いた東京大学生産技術研究所 坂内教授，東京大学空間情報科学研究センター岡部教授，杉盛助手に感謝いたします．

参考文献

[1]Open GIS Consortium, Inc., “OpenGIS Simple Features Specification for SQL Revision 1.0”, http://www.opengis.org/public/sfrl/sfsql_rev_1_0.pdf

[2]三浦信幸，横路誠司，高橋克巳，島健一，“GIS を用いた位置指向の WWW サーチエンジン～モーバイルインフォ 2 実験～”，地理情報システム学会講演論文集，Vol.7，pp.131-136，1998

[3]松本裕治，北内啓，山下達雄，今一修，今村友明，“日本語形態素解析システム『茶筌』version 1.0 使用説明書”，NAIST Technical Report，NAIST-IS-TR97007，February 1997，http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[4]XML/SGML サロン，“標準 XML 完全解説”，技術評論社，1998

[5]政木仁，“全国温泉案内『Ｏｈ！ＹＵ』”，http://www.csn.co.jp/ohyuhm.htm