SpatioTemporal MITRE-Sponsored Research

SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language

October 1, 2007

Version 2.0

Contact: hitz@mitre.org

Acknowledgements

SpatialML 2.0 is the first release of the guidelines for marking up Spatial ML, a markup language developed under funding from the MITRE Technology Program. The following people contributed ideas towards the development of Version 2.0:

- Dave Anderson (MITRE)
- Jade Goldstein-Stewart (Department of Defense)
- Amal Fayad-Beidas (MITRE)
- Dave Harris (MITRE)
- Dulip Herath (University of Colombo)
- Qian Hu (MITRE)
- Janet Hitzeman (MITRE)
- Seok Bae Jang (Georgetown University)
- Inderjeet Mani (MITRE)
- James Pustejovsky (Brandeis University)
- Justin Richer (MITRE)

This version will be posted at:

http://www.macforge.com/projects.php?cat=133&view=extended&n=50&page=6

We expect that subsequent releases will incorporate feedback from many others in the research community.

1 Introduction

We have developed a rich markup language called SpatialML for spatial locations, allowing potentially better integration of text collections with resources such as databases that provide spatial information about a domain, including gazetteers, physical feature databases, mapping services, etc.

Our focus is primarily on geography and culturally-relevant landmarks, rather than biology, cosmology, geology, or other regions of the domain of spatial language. However, we expect that these guidelines could be adapted to other such domains with some extensions, without changing the fundamental framework.

Our guidelines indicate language-specific rules for marking up SpatialML tags in English, as well as language-independent rules for marking up semantic attributes of tags. A handful of multilingual examples are provided in Section 16.

The main SpatialML tag is the PLACE tag. The central goal of SpatialML is to map PLACE information in text to data from gazetteers and other databases to the extent possible. Therefore, semantic attributes such as country abbreviations, country subdivision and dependent area abbreviations (e.g., US states), and geo-coordinates are used to help establish such a mapping. LINK and PATH tags express relations between places, such as inclusion relations and trajectories of various kinds. Information in the tag along with the tagged location string should be sufficient to uniquely determine the mapping, when such a mapping is possible. This also means that we don't include redundant information in the tag.

In order to make SpatialML easy to annotate without considerable training, the annotation scheme is kept fairly simple, with straightforward rules for what to mark and with a relatively "flat" annotation scheme. Further lightening is also possible, as indicated in Section 22.

2 Building on Prior Work

The goal in creating this spatial annotation scheme is to emulate the progress made earlier on time expressions, where the TIMEX2 annotation scheme for marking up such expressions[1] was developed and used in various projects for different languages, as well as schemes for marking up events and linking them to times, e.g., TimeML temporal linking[2] and the 2005 Automatic Content Extraction (ACE) guidelines.[3]

To the extent possible, SpatialML leverages ISO and other standards towards the goal of making the scheme compatible with existing and future corpora. The SpatialML guidelines are compatible with existing guidelines for spatial annotation and existing corpora within the ACE research program. In particular, we exploit the English Annotation Guidelines for Entities (Version 5.6.6 2006.08.01), specifically the GPE, Location, and Facility entity tags, and the Physical relation tags, all of which are mapped to SpatialML tags. We also borrow ideas from Toponym Resolution Markup Language of Leidner (2006), the research of Schilder et al. (2004) and the annotation scheme in Garbin and Mani (2005). Information recorded in the annotation is compatible with the feature types in the Alexandria Digital Library.[4] We also leverage the integrated gazetteer database (IGDB) of (Mardis and Burger 2005). Last but not least, this

[1]http://timex2.mitre.org
[2]http://www.timeml.org
[3]http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html
[4]http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/top.htm.

annotation scheme can be related to the Geography Markup Language (GML)[5] defined by the Open Geospatial Consortium (OGC), as well as Google Earth's Keyhole Markup Language (KML)[6] to express geographical features.

Our work goes beyond these schemes, however, in terms of providing a richer markup for natural language that includes semantic features and relationships that allow mapping to existing resources such as gazetteers. Such a markup can be useful for (i) disambiguation (ii) integration with mapping services, and (iii) spatial reasoning. In relation to (iii), it is possible to use spatial reasoning not only for integration with applications, but for better information extraction, e.g., for disambiguating a place name based on the locations of other place names in the document. We go to some length to represent topological relationships among places, derived from the RCC8 Calculus (Randell et al. 1992, Cohn et al. 1997).

The initial version of this annotation scheme focuses on toponyms and relative locations. In these examples, codes and special symbols can be found in the tables throughout the paper and those in Chapter 13. The least obvious of the codes will be listed near the examples. Geo-coordinates or gazetteer unique identifiers will be provided on occasion, but in general it is far too onerous to include them for each example in the guidelines.

3 Extent Rules (English-specific)

The rules for which PLACEs should be tagged are kept as simple as possible:

- Essentially, we tag any expression as a PLACE if it refers to a TYPE found in Table 4 (such as COUNTRY, STATE and RIVER). Do not mark phrase such as "here" or "the school" or "the Post Office."
- PLACEs can be in the form of proper names ("New York") or nominals ("town"), i.e. NAM or NOM.
- Adjectival forms of proper names ("U.S.," "Brazilian") are, however, tagged in order to allow us link expressions such as "Georgian" to "capital" in the phrase "the Georgian capital."[7]
- Non-referring expressions, such as "city" in "the city of Baton Rouge" are NOT tagged; their use is simply to indicate a property of the PLACE, as in this case, indicating that Baton Rouge is a city. In contrast, when "city" does refer, as in "John lives in the city" where "the city," in context, must be interpreted as referring to Baton Rouge, it is tagged as a place and given the coordinates, etc., of Baton Rouge.

[5] http://www.opengis.net/gml/
[6] http://earth.google.com/kml/
[7] This choice forces us to tag non-referring proper names in expressions such as "the non-U.S. team." The nonLocUse attribute on the PLACE tag is set to "true" in these cases.

- In general, extents of places which aren't referring expressions aren't marked, e.g., we won't mark any items in "a small town is better to live in than a big city."

The rules for what span ('extent') of text to mark for a PLACE are also kept as simple as possible:

- Premodifiers such as adjectives, determiners, etc. are NOT included in the extent unless they are part of a proper name. For example, for "the river Thames," only "Thames" is marked, but, for the proper names "River Thames" and "the Netherlands," the entire phrase is marked.
- Essentially, we try to keep the extents as small as possible, to make annotation easier.
- We see no need for tag embedding, since we have non-consuming tags (LINK and PATH) to express relationships between PLACEs.
- In the corpus we are releasing, we do NOT tag FACILITIES. The tagging of facilities is expected to be application-dependant.

4 Toponyms

Toponyms are proper names for places, and constitute a proper subset of the spatial locations described by SpatialML. We use a classification which allows most of the toponyms to be easily mapped to geo-coordinates (points or polygons) via a gazetteer. The classes are consolidated from two gazetteers: the USGS GNIS gazetteer and the NGA gazetteer. The Geographic Names Information System (GNIS), developed by the U.S. Geological Survey in cooperation with the U.S. Board on Geographic Names, contains information about physical and cultural geographic features in the United States and associated areas, both current and historical (not including roads and highways).[8] The National Geospatial-Intelligence Agency (NGA) gazetteer is a database of foreign geographic feature names with world-wide coverage, excluding the United States and Antarctica.[9] The consolidation is done in the IGDB gazetteer (Mardis and Burger 2005) developed at MITRE for the Disruptive Technologies Office.

4.1 Mapping Continents, Countries, and Country Capitals

The values COUNTRY, CONTINENT, and PPLC for the type feature are sufficient to disambiguate the corresponding PLACEs. There is no real need to add in geo-coordinates, since the latter can be determined unambiguously from a gazetteer. However, a gazetteer may be needed to establish that a place name is in fact the name of a country or capital.

[8] http://nhd.usgs.gov/gnis.html
[9] http://gnswww.nga.mil/geonames/GNS/index.jsp

Country codes are ISO-3166-1 two-letter codes. For countries not in ISO-3166-1, Yugoslavia, Czechoslovakia, Soviet Union, etc.), use the code OTHER.

Note: In these guidelines, we offer examples consisting of text paired with markup. In the text, all the SpatialML expressions being annotated are indicated with brackets, and below each example the corresponding markup is shown.

[Mexico] is in [North America]

<PLACE type="COUNTRY" country="MX" form="NAM">Mexico</PLACE>

<PLACE type="CONTINENT" continent="NA" form="NAM">North America</PLACE>

I attended a pro-[Iraqi] rally

<PLACE type="COUNTRY" country="IQ" form="NAM">Iraqi</PLACE>

The rest of [America] voted for Gore.

<PLACE type="COUNTRY" country="US" form="NAM">America</PLACE>

I rooted for the [US] team, even though Pele was playing on the [Brazilian] side.

<PLACE type="COUNTRY" country="US" form="NAM">US</PLACE>

<PLACE type="COUNTRY" country="BR" form="NAM">Brazilian</PLACE>

I visited many trattorias in [Rome], [Italy]

<PLACE type="PPLC" country="IT" form="NAM">Rome</PLACE>

<PLACE type="COUNTRY" country="IT" form="NAM">Italy</PLACE>

Table 1, below, shows the codes for the feature country, based on ISO-3166-1. Of course, there have been and will be countries not in Table 1. ISO-3166-2 is used for provinces. Because the standards are periodically updated, some oddities may arise; for example, as we write this document the country code for Hong Kong is HK (ISO-3166-1) but Hong Kong is also given a

province code of CN-91 (ISO-3166-2).[10]

[10] Similarly, Macao is listed as the province CN-92 and Taiwan is CN-71 in ISO-3166-2, while they also have country codes in ISO-3166-1.

AFGHANISTAN	AF	LIBERIA	LR
ÅLAND ISLANDS	AX	LIBYAN ARAB JAMAHIRIYA	LY
ALBANIA	AL	LIECHTENSTEIN	LI
ALGERIA	DZ	LITHUANIA	LT
AMERICAN SAMOA	AS	LUXEMBOURG	LU
ANDORRA	AD	MACAO	MO
ANGOLA	AO	MACEDONIA, THE FORMER YUGOSLAV REPUBLIC OF	MK
ANGUILLA	AI	MADAGASCAR	MG
ANTARCTICA	AQ	MALAWI	MW
ANTIGUA AND BARBUDA	AG	MALAYSIA	MY
ARGENTINA	AR	MALDIVES	MV
ARMENIA	AM	MALI	ML
ARUBA	AW	MALTA	MT
AUSTRALIA	AU	MARSHALL ISLANDS	MH
AUSTRIA	AT	MARTINIQUE	MQ
AZERBAIJAN	AZ	MAURITANIA	MR
BAHAMAS	BS	MAURITIUS	MU
BAHRAIN	BH	MAYOTTE	YT
BANGLADESH	BD	MEXICO	MX
BARBADOS	BB	MICRONESIA, FEDERATED STATES OF	FM
BELARUS	BY	MOLDOVA, REPUBLIC OF	MD
BELGIUM	BE	MONACO	MC
BELIZE	BZ	MONGOLIA	MN
BENIN	BJ	MONTENEGRO	ME
BERMUDA	BM	MONTSERRAT	MS
BHUTAN	BT	MOROCCO	MA
BOLIVIA	BO	MOZAMBIQUE	MZ
BOSNIA AND HERZEGOVINA	BA	MYANMAR	MM
BOTSWANA	BW	NAMIBIA	NA
BOUVET ISLAND	BV	NAURU	NR
BRAZIL	BR	NEPAL	NP
BRITISH INDIAN OCEAN TERRITORY	IO	NETHERLANDS	NL
BRUNEI DARUSSALAM	BN	NETHERLANDS ANTILLES	AN
BULGARIA	BG	NEW CALEDONIA	NC
BURKINA FASO	BF	NEW ZEALAND	NZ

BURUNDI	BI	NICARAGUA	NI
CAMBODIA	KH	NIGER	NE
CAMEROON	CM	NIGERIA	NG
CANADA	CA	NIUE	NU
CAPE VERDE	CV	NORFOLK ISLAND	NF
CAYMAN ISLANDS	KY	NORTHERN MARIANA ISLANDS	MP
CENTRAL AFRICAN REPUBLIC	CF	NORWAY	NO
CHAD	TD	OMAN	OM
CHILE	CL	PAKISTAN	PK
CHINA	CN	PALAU	PW
CHRISTMAS ISLAND	CX	PALESTINIAN TERRITORY, OCCUPIED	PS
COCOS (KEELING) ISLANDS	CC	PANAMA	PA
COLOMBIA	CO	PAPUA NEW GUINEA	PG
COMOROS	KM	PARAGUAY	PY
CONGO	CG	PERU	PE
CONGO, THE DEMOCRATIC REPUBLIC OF THE	CD	PHILIPPINES	PH
COOK ISLANDS	CK	PITCAIRN	PN
COSTA RICA	CR	POLAND	PL
CÔTE D'IVOIRE	CI	PORTUGAL	PT
CROATIA	HR	PUERTO RICO	PR
CUBA	CU	QATAR	QA
CYPRUS	CY	RÉUNION	RE
CZECH REPUBLIC	CZ	ROMANIA	RO
DENMARK	DK	RUSSIAN FEDERATION	RU
DJIBOUTI	DJ	RWANDA	RW
DOMINICA	DM	SAINT HELENA	SH
DOMINICAN REPUBLIC	DO	SAINT KITTS AND NEVIS	KN
ECUADOR	EC	SAINT LUCIA	LC
EGYPT	EG	SAINT PIERRE AND MIQUELON	PM
EL SALVADOR	SV	SAINT VINCENT AND THE GRENADINES	VC
EQUATORIAL GUINEA	GQ	SAMOA	WS
ERITREA	ER	SAN MARINO	SM
ESTONIA	EE	SAO TOME AND PRINCIPE	ST
ETHIOPIA	ET	SAUDI ARABIA	SA
FALKLAND ISLANDS (MALVINAS)	FK	SENEGAL	SN

FAROE ISLANDS	FO	SERBIA	RS
FIJI	FJ	SEYCHELLES	SC
FINLAND	FI	SIERRA LEONE	SL
FRANCE	FR	SINGAPORE	SG
FRENCH GUIANA	GF	SLOVAKIA	SK
FRENCH POLYNESIA	PF	SLOVENIA	SI
FRENCH SOUTHERN TERRITORIES	TF	SOLOMON ISLANDS	SB
GABON	GA	SOMALIA	SO
GAMBIA	GM	SOUTH AFRICA	ZA
GEORGIA	GE	SOUTH GEORGIA AND THE SOUTH SANDWICH ISLANDS	GS
GERMANY	DE	SPAIN	ES
GHANA	GH	SRI LANKA	LK
GIBRALTAR	GI	SUDAN	SD
GREECE	GR	SURINAME	SR
GREENLAND	GL	SVALBARD AND JAN MAYEN	SJ
GRENADA	GD	SWAZILAND	SZ
GUADELOUPE	GP	SWEDEN	SE
GUAM	GU	SWITZERLAND	CH
GUATEMALA	GT	SYRIAN ARAB REPUBLIC	SY
GUERNSEY	GG	TAIWAN, PROVINCE OF CHINA	TW
GINEA	GN	TAJIKISTAN	TJ
GUINEA-BISSAU	GW	TANZANIA, UNITED REPUBLIC OF	TZ
GUYANA	GY	THAILAND	TH
HAITI	HT	TIMOR-LESTE	TL
HEARD ISLAND AND MCDONALD ISLANDS	HM	TOGO	TG
HOLY SEE (VATICAN CITY STATE)	VA	TOKELAU	TK
HONDURAS	HN	TONGA	TO
HONG KONG	HK	TRINIDAD AND TOBAGO	TT
HUNGARY	HU	TUNISIA	TN
ICELAND	IS	TURKEY	TR
INDIA	IN	TURKMENISTAN	TM
INDONESIA	ID	TURKS AND CAICOS ISLANDS	TC
IRAN, ISLAMIC REPUBLIC OF	IR	TUVALU	TV
IRAQ	IQ	UGANDA	UG

IRELAND	IE	UKRAINE	UA
ISLE OF MAN	IM	UNITED ARAB EMIRATES	AE
ISRAEL	IL	UNITED KINGDOM	GB
ITALY	IT	UNITED STATES	US
JAMAICA	JM	UNITED STATES MINOR OUTLYING ISLANDS	UM
JAPAN	JP	URUGUAY	UY
JERSEY	JE	UZBEKISTAN	UZ
JORDAN	JO	VANUATU	VU
KAZAKHSTAN	KZ	Vatican City State see HOLY SEE	
KENYA	KE	VENEZUELA	VE
KIRIBATI	KI	VIETNAM	VN
KOREA, DEMOCRATIC PEOPLE'S REPUBLIC OF	KP	VIRGIN ISLANDS, BRITISH	VG
KOREA, REPUBLIC OF	KR	VIRGIN ISLANDS, U.S.	VI
KUWAIT	KW	WALLIS AND FUTUNA	WF
KYRGYZSTAN	KG	WESTERN SAHARA	EH
LAO PEOPLE'S DEMOCRATIC REPUBLIC	LA	YEMEN	YE
LATVIA	LV	Zaire	see CONGO, THE DEMOCRATIC REPUBLIC OF THE
LEBANON	LB	ZAMBIA	ZM
LESOTHO	LS	ZIMBABWE	ZW

Table 1: Country Codes (From ISO-3166 at http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html)

Table 2 shows the codes for continents:

AF	Africa
AN	Antarctica
AI	Asia
AU	Australia
EU	Europe
GO	Gondwanaland
LA	Laurasia
NA	North America

PA	Pangea
SA	South America

Table 2: Continent Codes (ca. 2000 A.E.)

4.2 Mapping via Gazetteer Unique Identifiers

Many place names are not of type COUNTRY, CONTINENT, and PPLC. For these, we map them if possible to a gazetteer reference. In the following example, "Madras" is a toponym and mappable by an annotator. To indicate the mapping, we use a unique identifier in the IGDB gazetteer via the gazref feature. Any authoritative gazetteer can be used, provided the gazetteer name is prefixed to the unique identifier.

The city of [Madras] is in a garrulous, Tamil-speaking [area].

<PLACE id=1 type="PPLA" country="IN"

form="NAM" gazref="IGDB:17896959">Madras</PLACE>

<PLACE id= 2 type="RGN" country="IN" form="NOM">area</PLACE>

<LINK source=2 target=1 linkType="IN">

(The form attribute and LINK tags will be explained below.)

Some places can be disambiguated but aren't construed as points that can be represented by pairs of geo-coordinates. Such places require polygons or other shapes to be characterized precisely. Providing gazetteer ids (via the gazref feature) is ideal for such cases, as the actual geometric description may be retrieved if needed offline. Some examples:

He cruised down the [Danube].

<PLACE type="WATER" form="NAM"

gazref="IGDB:209130408">Danube</PLACE>

He is an expert on [Himalayan] wildflowers.

<PLACE type="MTS" gazref="IGDB:209169910">Himalayan</PLACE>

n general, it is preferable to use a reliable gazetteer gazref to a latLong as the former provides evidence for the geo-coordinate that the gazref maps to.

The gazref is of the form **<gazetteer>:<gazid>**. It is allowable to use more than one gazetteer for providing gazrefs; It may be useful to use a different gazetteer when the primary gazetteer doesn't contain the place to be tagged.

4.3 Mapping via Geo-Coordinates

Sometimes the appropriate unique identifier will map to a gazetteer entry that lacks a geo-coordinate for some reason. Large bodies of land such as countries and continents, for example, will not have latitude/longitude information. In these cases, the gazref is still useful because an entry in a gazetteer may provide additional information about the PLACE, such as population or inclusion in other PLACEs.

If a gazetteer entry provides latitude/longitude information, we would include a geo-coordinate in the PLACE tag via the latLong feature.

Some places may not be present in a standard gazetteer at all, but may be provided with a geo-coordinate by some other method, such as using Google Earth or WordNet:

<PLACE type="FAC" id=3 form="NAM" gazref="GoogleEarth:xxxx"
latLong="40.45N 73.59W" description="great place to shop">Macy's</PLACE>

Geo-coordinates are to be used only for places that can be construed as points. Of course, a point given by a pair of geo-coordinates based on a reference coordinate system is at best an abstraction at some level of resolution. Here is an example of a typical geo-coordinate reference:

When walking in [New York City], watch out for dog-droppings.

<PLACE type="PPL" state="NY" country="US" latLong="40.714N 74.006W"
form="NAM">New York City</PLACE>

We allow the latLong feature to be any string, including strings with or without decimals that can be parsed into GML coordinates along with appropriate coordinate systems, including military coordinate systems. The Section below on GML mapping describes how to specify more meta-information about the geo-coordinate.

4.4 UnMappable Places

Sometimes it will not be possible for a human to extract a feature description for a toponym from the text, not even an ambiguous or abstract one. Examples include cases where the region has a non-standard boundary, such as "the Middle East." In such cases, it is still worthwhile to annotate whatever information can be gleaned from the text in the event that the gazetteer in question gets expanded in the future. SpatialML here offers only a little more information than ACE provides, without guaranteeing an ability to find a useful reference to the location in terms of a gazetteer. In such cases, using a gazetteer during annotation may not be helpful.

a bride from the [Middle East]

<PLACE type="RGN" form="NAM">Middle East</PLACE>

while traveling in the southern [Caucasus]

<PLACE type="RGN" mod="S" form="NAM">Caucasus</PLACE>

It is worth noting, however, that sometimes phrases of this type can be found in gazetteers. The IGDB, for example, has an entry for "Southwest," meaning the southwestern area of the United States. It doesn't hurt to look.

Gazetteers aren't perfect; there will be missing or inaccurate information in the gazetteer. Thus, a feature description may be of the kind which could refer to a gazetteer entry, but the entry may not be there, or it may be entered with the wrong geo-coordinates. In the former case, the annotator simply tags the location in the text without the gazetteer information. In the latter case, the annotator can ignore the gazetteer information if she knows it to be incorrect.

Dave is from [Tonawanda], not typically found in certain gazetteers.

<PLACE form="NAM">Tonawanda</PLACE>

5 Ambiguity in Mapping

5.1 Ambiguity in Text

It may often be the case that the text doesn't provide enough information for the human to map it to a unique geographical entry. In the following example, "Rochester" may refer to the city in Illinois or the one in New York State:

He arrived, in a vegetative state, in [Rochester].

If the text is genuinely ambiguous, we tag the place without any gazetteer reference or geo-coordinate.

<PLACE country="US" form="NAM">Rochester</PLACE>

5.2 Genuine Ambiguity in Gazetteer

In other cases, the text may make it clear which place is intended, at a level of granularity sufficient for understanding the text. However, such a level of granularity may be too coarse-grained compared to information found in the gazetteer:

He arrived, in a disturbed state, in [Rochester], [Illinois].

<PLACE id=1 state="IL" country="US" form="NAM">Rochester</PLACE>

<PLACE id=2 state="IL" country="US" type="STATE" form="NAM"
gazref="IGDB:30125575">Illinois</PLACE>

<LINK source=1 target=2 linkType="IN"/>

The feature description for *Rochester* yields three entries in USGS GNIS: one of type PPL (populated place) and one of type CIVIL (administrative area) in Sangamon county, Illinois with slightly different geo-coordinates (394458N 0893154W and 394446N 0893159W, respectively), and one of type PPL in Wabash county, Illinois with a different geo-coordinate (382044N 0874941W).

Clearly, we know that it's a Rochester in Illinois, but we don't know which county in Illinois is involved. Given the ambiguity, we have to leave out the gazref.

5.3 Multiple Gazetteer Entries for the Same Place

When there is more than one entry in the gazetteer for the same place, as one will often find in a gazetteer such as the IGDB which integrates several other gazetteers, prefer the entry which has a latlong over other entries. If ambiguity still remains, maintaining consistency of annotation is more complex. We recommend choosing the first entry that has a lat long, and, if none, then the first other entry that correctly maps the PLACE.[11]

[11] The IGDB contains many entries which are searchable under the form "X,Y" as in "Indiana, State of." These entries are likely to contain latlongs when the corresponding entry for the state name alone, "Indiana," does not. In order to test for these types of examples, it is worth trying the query "X,%" where % is a wildcard. The result will give latlongs for PLACEs such as "The Commonwealth of Massachusetts" and "The Kingdom of The Netherlands."

If the gazetteer supports equivalence class filtering, we leave out the gazref and the geo-coordinate.

5.4 When the Gazetteer is too Fine-Grained Compared to Text

Continuing the previous example, even if we know that Sangamon county is intended, we may not know which type of place Rochester should be.

He arrived, whining about the long bus ride, in the town of [Rochester], located in good old [Sangamon County], [Illinois].

Here we have a choice between a place of type PPL (with geo-coordinate 394458N 0893154W) and one of type CIVIL (with geo-coordinate 394446N 0893159W). Ambiguity of type being CIVIL or PPL is quite common, since towns and cities are not always marked in gazetteers as PPL, but are sometimes marked as CIVIL (an administrative region), reflecting the multiple views one can have of a place based on different criteria.

```
<PLACE id=1 state="IL" country="US" form="NAM" ctv="TOWN">Rochester</PLACE>
<PLACE id=2 state="IL" country="US" type="RGN"
        form="NAM">Sangamon County</PLACE>
<PLACE id=3 state="IL" country="US" type="STATE" form="NAM"
        gazref="IGDB:30125575">Illinois</PLACE>
<LINK source=1 target=2 linkType="IN"/>
<LINK source=2 target=3 linkType="IN"/>
```

Note: some gazetteer interfaces will support equivalence class filtering (as the IGDB interface does). Such a filter groups together all places that are treated as equivalent because they refer to the same place within some particular margin of error.

Note that mods never have a tagged extent.

6 Mapping Restrictions via the MOD attribute

Often the text will specify some restriction on the place. The MOD attribute is used to specify the type of restriction.

Fried okra is popular in the southern [United States]

<PLACE type="RGN" mod="S" country="US" form="NAM">United States</PLACE>

He mastered Swahili while living in [East Africa]

<PLACE type="RGN" mod="E" continent="AF" form="NAM">East Africa</PLACE>

Note that unlike "East Africa," "South Africa" is a proper name of a country, and providing its country code but no mod value is all that's needed to disambiguate it.

Table 3 shows the codes for mod. The types of mods are underlined, while the PLACEs are indicated in square brackets. Note that the mods are not tagged, just reflected in the value of the mod attribute in a PLACE tag. A mod phrase is only tagged if it is part of the PLACE name, as in the previous example.

B	the bottom of the [well]
BR	[Burmese] border
C	central [district]
E	eastern [province]
L	left on [Bourbon Street]
N	[North India]
NEAR	near [Harvard]
R	turn right at the [McDonald's]
S	southern [India]
T	the top of the [mountain]

W	west [Tikrit]

Table 3: MOD Codes

7 Using the Type Feature

It is crucial for an annotation scheme like SpatialML to provide a well-defined classification of places into different types that allow them to be mapped to geographical entries. However, there are several challenges in building such a typology:

- Too fine-grained a list of types (more than a dozen or so categories to choose from) will complicate the decision for human annotators. For machines, there are likely to be too few examples, and uneven distributions of examples for categories.
- Too coarse-grained a list of types may be of little use for a real application.
- Any such list is bound to be somewhat eclectic and application-driven.

We drew our types opportunistically from the NGA, USGS, and IGDB gazetteers. The Alexandria Digital Library (ADL) Feature Type Thesaurus, which the IGDB gazetteer is based on, classifies geographic entities into six top-level categories, with a further 205 categories below. The relevant fragment of the ADL Thesaurus that maps to our type codes is shown below (with our codes shown in uppercase).

administrative areas=RGN (sometimes)

. political areas

..countries=COUNTRY

..countries, 1st order divisions=CIVIL (sometimes)

..countries, 2nd order divisions=CIVIL (sometimes)

..countries, 3rd order divisions=CIVIL (sometimes)

..countries, 4th order divisions=CIVIL (sometimes)

.populated places=PPL, PPLA, PPLC, CIVIL

hydrographic features=WATER

manmade features=FAC

. transportation features

.. roadways=ROAD

physiographic features=RGN (sometimes)

.mountains=MTN

..mountain ranges=MTNS

regions=RGN (sometimes)

.land regions

..continents=CONTINENT

Table 4 shows the codes for type. This is by its very nature a partial list. The categories are mutually exclusive.

When the types CONTINENT, COUNTRY, STATE and LATLONG are chosen, the corresponding slots continent, country, state and latlong must be filled *only* if they are not specified by the gazref entry; to do so would be redundant. If the gazref entry does not contain a latlong, an attempt to find one should be made via Google, Wikipedia or elsewhere.

WATER	River, stream, ocean, sea, lake, canal, aqueduct, geyser, etc.
CELESTIAL	Sun, Moon, Jupiter, Gemini, etc.
CIVIL	Political Region or Administrative Area, usually sub-national, e.g. State, Province, certain instances of towns and cities.
CONTINENT	Denotes a continent, including ancient ones. See Table 2.
COUNTRY	Denotes a country, including ancient ones. See Table 1.
FAC	Facility, usually a catchall category for restaurants, churches, schools, ice-cream parlors, bowling alleys, you name it!
GRID	A grid reference indication of the location, e.g., MGRS (Military Grid Reference System)
LATLONG	A latitude/longitude indication of the location
MTN	Mountain
MTS	Range of mountains
POSTALCODE	Zip codes, postcodes, pin codes etc.
POSTBOX	P. O. Box segments of addresses
PPL	Populated Place (usually conceived of as a point), other than PPLA or PPLC
PPLA	Capital of a first-order administrative division, e.g., a state capital
PPLC	Capital of a country
RGN	Region other than Political/Administrative Region
ROAD	Street, road, highway, etc.
STATE	A first-order administrative division within a country, e.g., state, province, gubernia, territory, etc. See Table 7.

UTM	A Universal Transverse Mercator (UTM) format indication of the location
VEHICLE	Car, truck, train, etc.

Table 4: TYPE Codes

8 Annotating Text-Described Settlements with CTV

The commonsense notions of cities, towns and villages are particular types of settlements that are often hard to detect from gazetteer entries. We may be lucky and find a place to be of type PPLC, in which case we can determine it's a city. However, in other cases we may find it to be of type PPL and not know whether it's a city or town, or it may be of type CIVIL and be in fact a town or city.

We use the feature CTV (values CITY, TOWN, or VILLAGE) to annotate cases where the text explicitly specifies that a place is of that type. In these cases, the annotator should not guess, but use only the information made available by the text.

the town of [Rochester]

<PLACE ctv="TOWN" form="NAM">Rochester</PLACE>

9 Annotating Geo-Coordinates found in text

Some texts may contain geo-coordinates. Geo-coordinates found in texts may be ill-formed, incorrect, or in a different coordinate system from the gazetteer in use.

We distinguish between the geo-coordinate found in a text and one guaranteed to be well-formed by marking the former with a PLACE tag with a type value of LATLONG, GRID, or UTM, and placing the well-formed geo-coordinate in the latLong attribute of the PLACE. In the following example, a link of type EQ is required in order to indicate that the location of *Rochester* is the same as that of the latitude/longitude:

[Rochester], [Illinois] [394458N 0893154W]

<PLACE type="CITY" country="US" id=1 form="NAM">Rochester</PLACE>

<PLACE type="CIVIL" country="US" id=2 form="NAM">Illinois</PLACE>

<PLACE type="LATLONG" id=3>"394458N 0893154W"</PLACE>

<LINK id=4 source=1 target=2 linkType="IN"/>

<LINK id=5 source=1 target=3 linkType="EQ"/>

f the latLong value is taken from a gazetteer, the gazref attribute must also be given a value.

Once the string with the geo-coordinate is verified to be correct or is mapped onto the corresponding geo-coordinate type from a gazetteer, the resulting geo-coordinate is placed as the value of the PLACE latLong attribute, as below:

<PLACE type="LATLONG" id=3
latLong="394458N 0893155W">394458N 0893154W</PLACE>

10 Annotating Addresses

[100 James Drive, SE], [Vienna], [Virginia] [22180]
<PLACE type="ROAD">100 James Drive, SE</PLACE>
<PLACE type="PPL" state="VA">Vienna</PLACE>
<PLACE type="STATE" state="VA">Virginia</PLACE>
<PLACE type="POSTALCODE">22180</PLACE>

11 Marking Exceptional Information

Every tag has a comment attribute which can be used by the annotator to record difficulties in annotation. These should only be used in case of serious difficulty.

PLACE tags also have a nonLocUse feature. This is to be set to "true" for cases where the PLACE does not involve a location. Typically, this is a difficult decision to make, e.g., should *U.S.* in *the U.S. team* be marked as nonLocUse or not? To say yes in this case would revert back to the GPE/non-GPE distinction in ACE which caused the annotators difficulty. The nonLocUse feature is therefore to be used when the view of the place as a location corresponding to that mention would be entirely misleading, e.g., *non-U.S. interests*.

12 Annotating Relative Locations via Spatial Relations

12.1 PATHs

We use a PATH tag to express a spatial trajectory between a pair of locations. For example:

[Amritsar], [northwest] of the capital [New Delhi]

<PLACE type="PPL" id=1 form="NAM">Amritsar</PLACE>

<PLACE type="PPLC" country="IN" id=2 form="NAM">New Delhi</PLACE>

<SIGNAL id=3>northwest</SIGNAL>

<PATH direction="NW" source=2 destination=1 signals="3" />

The PATH indicates that in order to travel from source New Delhi to destination Amritsar you would go in the NW direction.

We also use SIGNAL tags to indicate the text portion that licenses the path. The SIGNAL should not include trailing prepositions, but each portion of the signal should be tagged individually, as in *[30 miles] [west] of the city*. Similarly, where the signals are discontinuous, they will be represented as multiple signals, e.g., *[two blocks down] and [one over] from the zoo.* The signal ids licensing the path may be included in a signals attribute in the path tag.

a [town] some [50 miles] [south] of [Salzburg] in the central [Austrian] [Alps]

<PLACE type="PPL" id=1 form="NOM" ctv="TOWN">town</PLACE>

<SIGNAL id=2>50 miles</SIGNAL>

<SIGNAL id=3>south</SIGNAL>

<PLACE id=4 type="PPLA" country="AU" form="NAM">Salzburg</PLACE>

<PLACE id=5 type="COUNTRY" country="AT" mod="C">Austrian</PLACE>

<PLACE id=6 type="MTS" >Alps</PLACE>

<PATH id=7 distance="50:mi" direction="S" source= 4 destination=1 signals="2 3"/>

<LINK id=8 source=1 target=6 linkType="IN"/>

<LINK id=9 source=6 target=5 linkType="IN"/>

Mark PATHs only when they are described within one phrase, i.e., if parts of a path are described in different sentences or in different parts of the same sentence, do not mark them.

For direction codes, refer to Table 5.

Direction	Example
B	[behind] the house
A	[above] the roof
BL	[below] the tree-line
E	[E] of
ESE, WSW, etc.	
F	[in front of] the theater
N	[north] of
S	[south] of
W	[W] of

Table 5: Codes for Directions

12.2 LINKs

We use a LINK tag to express containment, connection, or other topological relations between a pair of locations. Thus, in the above example, we use a linkType of IN (inclusion). Possible linkTypes are listed in Table 6. These are adapted from the RCC8 Calculus.

LinkType	Example
IN (tangential and non-tangential proper parts)	[Paris], [Texas]
EC (extended connection)	the border between [Lebanon] and [Israel]
NR (near)	visited [Belmont], near [San Mateo]
DC (discrete connection)	the [well] outside the [house]
PO (partial overlap)	[Russia] and [Asia]
EQ (equality)	[Rochester] and [382044N 0874941W]

Table 6: Codes for Link Types (partially derived from RCC8 Calculus)

Here are other common examples of inclusion:

[Moscow], [Russia]

<PLACE type="PPLC" country="RU" id=1 form="NAM">Moscow</PLACE>

<PLACE type="COUNTRY" country="RU" id=2 form="NAM">Russia</PLACE>

<LINK source=1 target=2 linkType="IN"/>

the basketball [arena] of [Michigan State University]

<PLACE type="FAC" id=1 form="NOM">arena</PLACE>

<PLACE type="FAC" id=2 form="NAM">Michigan State University</PLACE>

<LINK source=1 target=2 linkType="IN"/>

a [well] in [West Tikrit]

<PLACE type="FAC" id=1 form="NOM">well</PLACE>

<PLACE type="CIVIL" mod="W" country="IQ" id=2 form="NAM">West Tikrit</PLACE>

<LINK source=1 target=2 linkType="IN"/>

this northern [Uganda] [town]

<PLACE type="PPL" country="UG" ctv="TOWN" id=1>town</PLACE>

<PLACE type="COUNTRY" country="UG" mod="N" id=2>Uganda</PLACE>

<LINK source=1 target=2 linkType="IN"/>

the central [district] of the town of [Tirunelveli], [Tamil Nadu] in southern [India]

<PLACE type="RGN" mod=C id=1 form="NOM">district</PLACE>

<PLACE type="PPL" id=2 form="NAM">Tirunelveli</PLACE>

<PLACE type="CIVIL" country="IN" id=3 form="NAM">Tamil Nadu</PLACE>

<PLACE type="COUNTRY" country="IN" mod="S" id=4 form="NAM">India</PLACE>

<LINK source=1 target=2 linkType="IN"/>

<LINK source=2 target=3 linkType="IN"/>

To help determine the location of a place mentioned in the text, the entire document can be used as context by the annotator.

<LINK source=2 target=4 linkType="IN"/>

the hot dog [stand] [behind] the [Macy's] on [Broadway]

<PLACE type="FAC" id=1 form="NOM">stand</PLACE>

<SIGNAL id=2>behind</SIGNAL>

<PLACE type="FAC" id=3 form="NAM" gazref="GoogleEarth:xxxx"
latLong="40.45N 73.59W">Macy's</PLACE>

<PLACE type="ROAD" id=4 form="NAM">Broadway</PLACE>

<PATH id=5 direction="B" source=1 destination=3 frame="VIEWER" signals=2/>

<LINK source=3 target=4 linkType="IN"/>

[towards] [Scammonden Water] [along] the [B6114]

<SIGNAL id=1>towards</SIGNAL>

<PLACE type="WATER" country=UK id=2
form="NAM">Scammonden Water</PLACE>

<SIGNAL id=3>along</SIGNAL>

<PLACE type="ROAD" country="UK" id=4 form="NAM">B6114</PLACE>

<PATH id=5 destination=2 frame="VIEWER" signals="1 3"/>

<LINK id=6 source=5 target=4 linkType="EC">

The PATH tag in the above example indicates a path towards a destination (i.e., a body of water). The source is not specified. The LINK tag indicates that the path has an Extended Connection (EC) with (i.e., is running along) a road, via the use of the PATH id as the source of the LINK.

13 Disambiguation Guidelines

The annotator is not to use specialized knowledge that is not part of commonsense knowledge that everyone is expected to have.

ate codes are ISO-3166-2 codes. For states not in ISO-3166-2, ??????????????????????

Thus, given a bare mention of *Rome*, the annotator can use information from the entire document to determine which of the various places named "Rome" it is.

For example, if the text mentions a pizza joint in Rome, but doesn't otherwise specify which Rome it is, and if the pizza joint's description exactly matches the annotator's memory of a particular pizza joint allowing the annotator to identify which Rome it is, the annotator is not to indicate the correct Rome based on this knowledge. This issue may arise in certain texts such as the annotation of travel blogs, when the annotator has visited the location under discussion. The annotator must rely solely on the information in the text and in the gazetteer in order to keep the annotation more representative of general geospatial knowledge, and therefore more consistent with the work of other annotators.

14 States

States are top-level administrative divisions of countries. Like towns, cities and villages, they are an intuitive category that corresponds to different types of entities in gazetteers. State codes are ISO-3166-2 codes (excluding the country code and hyphen) (see http://en.wikipedia.org/wiki/ISO_3166-2)

Table 7 provides a list of state codes for US states..

AL	Alabama	KY	Kentucky	ND	North Dakota
AK	Alaska	LA	Louisiana	OH	Ohio
AZ	Arizona	ME	Maine	OK	Oklahoma
AR	Arkansas	MD	Maryland	OR	Oregon
CA	California	MA	Massachusetts	PA	Pennsylvania
CO	Colorado	MI	Michigan	RI	Rhode Island

CT	Connecticut	MN	Minnesota	SC	South Carolina
DE	Delaware	MS	Mississippi	SD	South Dakota
DC	District of Columbia	MO	Missouri	TN	Tennessee
FL	Florida	MT	Montana	TX	Texas
GA	Georgia	NE	Nebraska	UT	Utah
HI	Hawaii	NV	Nevada	VT	Vermont
ID	Idaho	NH	New Hampshire	VA	Virginia
IL	Illinois	NJ	New Jersey	WA	Washington
IN	Indiana	NM	New Mexico	WV	West Virginia
IA	Iowa	NY	New York	WI	Wisconsin
KS	Kansas	NC	North Carolina	WY	Wyoming

Table 7: Codes for US States

15 Inventory of SpatialML Tags

The full XML DTD for SpatialML is given at the end of the document. In Table 8, we list the tag attributes with some documentation. Each of these tags also has a comment field, as described in Section 11.

PLACE	county	When provided by the text
	state	From Table 11 or use non-US state abbreviation
	country	See Table 1
	continent	See Table 2
	ctv	CITY, TOWN, or VILLAGE (when indicated as such in the text)
	gazref	Single gazetteer id, e.g., IGDB. Prefix the id with the gazetteer name plus a colon, e.g., WordNet:310975, IGDB:2104656
	id	tagid
	latLong	When gazref is available, the coordinate from the gazetteer may be copied here
	mod	See Table 3
	type	See Table 4
	form	NAM (proper noun) or NOM (nominal)
	nonLocUse	e.g., “non-U.S. organizations”

	description	For a convenient textual description of the place found in the local context of the mention. This is intended for use by applications which provide their own criteria for how to fill the slot.
	comment	text field
PATH	source	tagid
	id	tagid
	destination	tagid
	direction	See Table 5
	distance	number:units
	frame	viewer, intrinsic, extrinsic
	signals	a string containing a list of tagids separated by a space
	comment	text field
LINK	source	tagid
	id	tagid
	target	tagid
	linkType	See Table 6
	comment	text field
SIGNAL	id	tagid
	comment	text field

Table 8: SpatialML Tags and Attributes

16 Multilingual Examples

SpatialML is intended as a language-independent markup language. Of course, the rules for what extents to mark may have to be adjusted based on the morphology and orthography of a particular language. In what follows, we present sentences from English, Arabic, Korean and Sinhala annotated in SpatialML. These are merely illustrative of the scope of SpatialML, and do not pretend to cover any idiosyncrasies in these languages in the way they talk about space. Further work on Mandarin is ongoing. Of course, more detailed investigation of spatial expressions in these languages would require a separate research effort.

1. I attended a pro-[American] rally.
<PLACE type="COUNTRY" country="US" form="NAM">American</PLACE>

Here is the corresponding Arabic.

حضرت مظاهرة مريدة للولا يات المتحدة
<PLACE type="COUNTRY" country="US" form="NAM">الولايات المتحدة </PLACE>

Turning to Korean:

나는 프로-[아메리칸] 랠리에 참가하였다.
I-Top pro-American rally-Loc attend-Past-ending
<PLACE type="COUNTRY" country="US" form="NAM">아메리칸</PLACE>

Note that both the English and Korean use sub-word tags.

Here is the corresponding Sinhala:

□□ *[□□□□□□□□□]*- □□□□□□□□ □□□□□□□□ □□□□□□□ □□□□.
<PLACE type=COUNTRY country=US form=NAM> □□□□□□□□□ </PLACE>

Now for the Mandarin:

*我*出席了*一个拥护[美国]的集会。*
<PLACE type="COUNTRY" country="US" form="NAM"> *美国*</PLACE>

2. I live in this northern [Uganda] [town].
<PLACE type="PPL" country="UG" id=1 ctv="TOWN">town</PLACE>
<PLACE type="COUNTRY" country="UG" mod="N" id=2>Uganda</PLACE>
<LINK source=1 target=2 linkType="IN"/>

أنا أسكن فى مدينة شمال أوغندا
<PLACE type="PPL" country="UG" id=1> مدينة</PLACE>
<PLACE type="COUNTRY" country="UG" mod="N" id=2> أوغندا</PLACE>
<LINK source=1 target=2 linkType="IN"/>

나는 이 [우간다] 북쪽 [마을]에 산다.
I-Top this [Uganda] northern [town]-Loc live-Present-ending.
<PLACE type="COUNTRY" country="UG" mod="N" id=1>우간다</PLACE>
<PLACE type="PPL" country="UG" id=2 ctv="TOWN">마을</PLACE>
<LINK source=2 target=1 linkType="IN"/>

Since the Korean word-order is different, the tag ids have changed slightly, but this difference is inconsequential.

□□ □□ □□□□□□ *[□□□□□□□] [□□□□□□□□□] □□□□□□□.*
□□□□ = *the town*
□□□□□□□ = *a town*
□□□□□□□□ = *in the town*
□□□□□□ = *in a town*
<PLACE type=PPL country=UG id=1 ctv=TOWN> □□□□□□□□ </PLACE>
<PLACE type=country country=UG mod=N id=2> □□□□□□□ </PLACE>
<LINK source=1 target=2 linkType=IN>

我居住在这个北[乌干达] [镇]。

<PLACE type="PPL" country="UG" id=1 ctv="TOWN">*镇*</PLACE>
<PLACE type="COUNTRY" country="UG" mod="N" id=2>*乌干达*</PLACE>
<LINK source=1 target=2 linkType="IN"/>

3. I live in [Amritsar], [northwest] of the capital [New Delhi].

<PLACE type="PPL" id=1 country="IN" form="NAM">Amritsar</PLACE>
<PLACE type="PPLC" country="IN" id=2 form="NAM">New Delhi</PLACE>
<SIGNAL id=3>northwest</SIGNAL>
<PATH direction="NW" source=2 destination=1 signals="3" />

أنا أسكن فى اميرستارشمال غرب العاصمة نيودلهى
<PLACE type="PPL" id=1 country="IN" form="NAM">اميرستار</PLACE>
<PLACE type="PPLC" country="IN" id=2 form="NAM">نيودلهى </PLACE>
<SIGNAL id=3>شمال غرب </SIGNAL>
<PATH direction="NW" source=2 destination=1 signals="3"/>

나는 수도 [뉴델리] [북서쪽]의 [암리차르]에 산다.
I-Top capital [New Delhi] [northwest] -Pos [Amritsar]-Loc live-Present-ending
<PLACE type="PPLC" country="IN" id=1 form="NAM">뉴델리</PLACE>
<SIGNAL id=2>북서쪽</SIGNAL>
<PLACE type="PPL" id=3 country="IN" form="NAM">암리차르</PLACE>
<PATH direction="NW" source=1 destination=3 signals="2" />

□□ *[□□ □□□□□□□] □□□□□□□□ [□□□□□□□] □□□□□□□ [□□□□□□□□□□□□□]*
□□□□□□□.
<PLACE type=PPL id=1 country=IN form=NAM> □□□□□□□□□□□□ </PLACE>

<PLACE type=PPLC country=IN id=2 form=NAM> □□ □□□□□□ </PLACE>
<SIGNAL id=3> □□□□□□ </SIGNAL>
<PATH direction=NW source=2 destination=1 signals=3 />

我住在[阿姆利则],在首都[新德里]的[西北部]。

<PLACE type="PPL" id=1 country="IN" form="NAM"> *阿姆利则* </PLACE>
<PLACE type="PPLC" country="IN" id=2 form="NAM"> *新德里* </PLACE>
<SIGNAL id=3>*西北部* </SIGNAL>
<PATH direction="NW" source=2 destination=1 signals="3" />

4. I live in a [town] some [50 miles] [south] of [Salzburg] in the central [Austrian] [Alps].
<PLACE type="PPL" id=1 form="NOM" ctv="TOWN">*town*</PLACE>
<SIGNAL id=2>*50 miles*</SIGNAL>
<SIGNAL id=3>*south*</SIGNAL>
<PLACE id=4 type="PPLA" country="AU" form="NAM">*Salzburg*</PLACE>
<PLACE id =5 type="COUNTRY" country="AU" mod=C>*Austrian*</PLACE>
<PLACE id =6 type="MTS" >*Alps*</PLACE>
<PATH id=7 distance="50:mi" direction=S source= 4 destination=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>

أنا أسكن فى مدينة تبعد حوالى خمسين ميل جنوب سالزبرج فى وسط النمسا و جبال الالب
<PLACE type="PPL" id=1 form="NOM"> مدينة</PLACE>
<SIGNAL id=2>خمسين ميل </SIGNAL>
<SIGNAL id=3>جنوب /SIGNAL>
<PLACE id=4 type="PPLA" country="AU" form="NAM">سالزبرج </PLACE>
<PLACE id=5 type="COUNTRY" country="AU" mod=C>النمسا <PLACE>
<PLACE id =6 type="MTS"> جبال الالب </PLACE>
<PATH id=7 distance="50:mi" direction=S source= 4 destination=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>

나는 [오스트리아] [알프스] 중심의 [잘츠부르크] [남쪽]에서 [50마일] 거리의 마을에 산다.
I-Top Austria Alps Center-Pos Salzburg south-From 50 miles distance-Pos town-Loc live-Present-ending
<PLACE type="PPL" id=1 form="NOM" ctv="TOWN">마을</PLACE>
<SIGNAL id=2>50 마일</SIGNAL>
<SIGNAL id=3>남쪽</SIGNAL>
<PLACE id=4.3 type="PPLA" country="AU" form="NAM">잘츠부르크</PLACE>

<PLACE id =4.1 type="COUNTRY" country=AT mod="C">오스트리아</PLACE>
<PLACE id =6.2 type="MTS" >알프스</PLACE>
<PATH id=7 distance="50:mi" direction=S source= 4 destination=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>

□□ □□□□□ *[□□□□□□□□□□□□] [□□□□□□□□] □□□□□□□ □□□□□□ [□□□□□□□□□□□□□□□□□□] [□□□□□□□ 50□□] □□□ [□□□□□□□□] □□□□□□ [□□□□□] □□□□□□*.
<PLACE type=PPL id=1 form=NOM ctv=TOWN> □□□□□ </PLACE>
<SIGNAL id=2> □□□□□□□ *50□□* </SIGNAL>
<SIGNAL id=3> □□□□□□□ </SIGNAL>
<PLACE id=4 type=PPLA country=AU form=NAM> □□□□□□□□□□□□□□□□□□
</PLACE>
<PLACE id =5 type=country country=AT mod=C> □□□□□□□□□□□□ </PLACE>
<PLACE id =6 type=MTS > □□□□□□□ </PLACE>
<PATH id=7 distance="50:mi" direction=S source= 4 destination=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType=IN/>

我居住在一个离中[奥地利] [阿尔卑斯] [萨尔茨堡] [以南]大约 [50 英哩] 的 [镇子]里。
<PLACE id =1 type="COUNTRY" country="AU" mod=C> 奥地利 </PLACE>
<PLACE id =2 type="MTS" >阿尔卑斯</PLACE>
<PLACE id=3 type="PPLA" country="AU" form="NAM"> 萨尔茨堡 </PLACE>
<SIGNAL id=4>以南</SIGNAL>
<SIGNAL id=5>50 英哩</SIGNAL>
<PLACE type="PPL" id=6 form="NOM" ctv="TOWN"> 镇子</PLACE>
<PATH id=7 distance="50:mi" direction=S source=3 destination=6 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>

5. I met Laila in a [cafe] in [Rabat].
<PLACE id=1 form="NOM" type=FAC>*cafe*</PLACE>
<PLACE type="PPLC" id=2 form="NAM" country="MA">*Rabat*</PLACE>
<LINK id=3 source=1 target=2 linkType="IN"/>

ألتقيت بليلى في مقهي في الرباط
<PLACE type="COUNTRY" id=1 form="NAM" country="IQ">مقهي </PLACE>
<PLACE type="PPLC" id=2 form="NAM" country="MA">الرباط </PLACE>
<LINK id=3 source=1 target=2 linkType="IN"/>

나는 [라바트]에 있는 [카페]에서 라일라를 만났다.
I-Top Rabat-Loc exist-ending cafe-Loc Laila-Acc meet-Past-ending

<PLACE id=1 form="NOM" type=FAC>카페</PLACE>
<PLACE type="PPLC" id=2 form="NAM" country="MA">라바트</PLACE>
<LINK id=3 source=1 target=2 linkType="IN"/>

□□ *[□□□□□□□] [□□□□□□□□] □□□□□ □□□□□□□□.*
<PLACE id=1 form=NOM type=FAC> □□□□□□□□ </PLACE>
<PLACE type=PPLC id=2 form=NAM country=MA> □□□□□□ </PLACE>
<LINK id=3 source=1 target=2 linkType=IN/>

我在[拉巴特]的一个[咖啡馆]遇见了莱拉。
<PLACE type="PPLC" id=1 form="NAM" country="MA"> *拉巴特* </PLACE>
<PLACE id=2 form="NOM" type=FAC>*咖啡馆*</PLACE>
<LINK id=3 source=2 target=1 linkType="IN"/>

6. I live in the key [Iraqi] border [town] of [Qaim].
<PLACE type="COUNTRY" id=1 form="NAM" country="IQ">*Iraqi*</PLACE>
<PLACE type="PPL" id=2 form="NOM" mod="BR" ctv="TOWN">*town*</PLACE>
<PLACE type="PPL" id=3 form="NAM" ctv="TOWN" country="IQ">*Qaim*</PLACE>
<LINK id=4 source=2 target=3 linkType="EQ"/>

أنا أسكن في المدينة العراقية الحدودية الرئيسية قم
<PLACE type="COUNTRY" id=1 form="NAM" country="IQ">المدينة العراقية </PLACE>
<PLACE type="PPL" id=2 form="NOM" mod="BR" ctv="TOWN"> المدينة </PLACE>
<PLACE type="PPL" id=3 form="NAM" ctv="TOWN" country="IQ"> قم </PLACE>
<LINK id=4 source=2 target=3 linkType="EQ"/>

나는 [콰임]의 이라크 [국경] 마을에 산다.
I-Top Qaim-Pos Iraq border town-Loc live-Present-ending

<PLACE type="COUNTRY" id=1 form="NAM" country="IQ">이라크</PLACE>
<PLACE type="PPL" id=2 form="NOM" mod="BR" ctv="TOWN">마을</PLACE>
<PLACE type="PPL" id=3 form="NAM" ctv="TOWN" country="IQ">콰임</PLACE>
<LINK id=4 source=2 target=3 linkType="EQ"/>

□□ □□□□□□ *[□□□□]* □□□□□□□□ □□□□ □□ *[□□□□□□□□□□□]* □□□□□□.
<PLACE type=country id=1 form=NAM country=IZ> □□□□ </PLACE>
<PLACE type=PPL id=2 form=NOM mod=BR ctv=TOWN> □□□□□ </PLACE>
<PLACE type=PPL id=3 form=NAM ctv=TOWN country=IZ> □□□□□□□□□□□ </PLACE>
<LINK id=4 source=2 target=3 linkType=EQ/>

我住在[伊拉克]边界的重[镇][奎坦]。

<PLACE type="COUNTRY" id=1 form="NAM" country="IQ"> *伊拉克*</PLACE>
<PLACE type="PPL" id=2 form="NOM" mod="BR" ctv="TOWN"> *镇*</PLACE>
<PLACE type="PPL" id=3 form="NAM" ctv="TOWN" country="IQ"> *奎坦*</PLACE>
<LINK id=4 source=2 target=3 linkType="EQ"/>

7. I was born in [Qaim], about [200 miles] [west] of [Baghdad].
<PLACE type="PPL" id=1 form="NAM" country="IQ">Qaim</PLACE>
<SIGNAL id=2>200 miles</SIGNAL>
<SIGNAL id=3>west</SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ">Baghdad</PLACE>
<PATH id=5 distance="200:mi" direction="W" source=1 destination=4 signals="2 3"/>

انا من مواليد مدينة قم حوالى مائتين ميلا غرب بغداد

<PLACE type="PPL" id=1 form="NAM" country="IQ"> قم <PLACE>
<SIGNAL id=2> مائتين ميلا </SIGNAL>
<SIGNAL id=3>غرب</SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ">بغداد</PLACE>
<PATH id=5 distance="200:mi" direction="W" source=1 destination=4 signals="2 3"/>

나는 [바그다드] [서쪽]으로 약 200 마일 거리의 [콰임]에서 태어났다.
I-Top Baghdad west-from about 200 mile distance-Pos Qaim-Loc born-Past-ending
<PLACE type="PPL" id=1 form="NAM" country="IQ">콰임</PLACE>
<SIGNAL id=2>200 마일</SIGNAL>
<SIGNAL id=3>서쪽</SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ">바그다드</PLACE>
<PATH id=5 distance="200:mi" direction="W" source=1 destination=4 signals="2 3"/>

□□ *[□□□□□□□□□□] [□□□□□□□□ 200□□]* □□□ *[□□□□□□□□□]* □□□□□□ *[□□□□□□□□□□□]* □□□□□□□□□.
<PLACE type=PPL id=1 form=NAM country=IZ> □□□□□□□□□□□ </PLACE>
<SIGNAL id=2> □□□□□□□□ *200□□* </SIGNAL>
<SIGNAL id=3> □□□□□□□□□ </SIGNAL>
<PLACE type=PPLC id=4 form=NAM country=IZ> □□□□□□□□□□ *</PLACE>*
<PATH id=5 distance="200:mi" direction=W source=1 destination=4 signals="2 3"/>

我出生在离[巴格达][西面]大约[二百英哩]的[奎坦]。

I was born in [Qaim], about [200 miles] [west] of [Baghdad].
<PLACE type="PPL" id=1 form="NAM" country="IQ"> *巴格达* </PLACE>
<SIGNAL id=2>*西面* </SIGNAL>
<SIGNAL id=3>*二百英里* </SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ"> *奎姆* </PLACE>
<PATH id=5 distance="200:mi" direction="W" source=1 destination=4 signals="2 3"/>

8. I live within [two miles] of the [Mexican] border.
<SIGNAL id=1>*two miles*</SIGNAL>
<PLACE type="COUNTRY" id=2 form="NAM" mod="BR"
country="MX">*Mexican*</PLACE>
<PATH id=3 distance="2:mi" source=2 signals="1"/>

أنا أسكن علي بعد ما يقارب من أثنين ميل من جدود المكسيك
<SIGNAL id=1> أثنين ميل </SIGNAL>
<PLACE type="COUNTRY" id=2 form="NAM" mod="BR"
country="MX"> المكسيك </PLACE>
<PATH id=3 distance="2:mi" source=2 signals="1"/>

나는 [멕시코] 국경에서 [2 마일] 안에 산다.
I-Top Mexico border-From 2 mile within-Loc live-Present-ending
<SIGNAL id=1>2 마일</SIGNAL>
<PLACE type="COUNTRY" id=2 form="NAM" mod="BR"
country="MX">멕시코</PLACE>
<PATH id=3 distance="2:mi" source=2 signals="1"/>

□□ *[□□□□□□□□□□□□]* □□□□□□□□□□ □□□ *[□□□□□□□□ 2□□]* □□□□□ □□□□□□.
<SIGNAL id=1>□□□□□□□□ *2□□* </SIGNAL>
<PLACE type=country id=2 form=NAM mod=BR country=MX>□□□□□□□□□□□
</PLACE>
<PATH id=3 distance="2:mi" source=2 signals="1"/>

我住在离[墨西哥]边境[两英哩]以内。
I live within [two miles] of the [Mexican] border.

<PLACE type="COUNTRY" id=1 form="NAM" mod="BR"
country="MX"> *墨西哥* </PLACE>
<SIGNAL id=2>*两英里* </SIGNAL>
<PATH id=3 distance="2:mi" source=1 signals="1"/>

9. I traveled [along] the [Euphrates River].
<SIGNAL id=1>along</SIGNAL>
<PLACE type="WATER" id=2 form="NAM">*Euphrates River*</PLACE>
<PATH id=3 frame="VIEWER" signals="1"/>
<LINK id=4 source=3 target=2 linkType="EC">

سافرت علي جانب نهرالفرات
<SIGNAL id=1> علي جانب </SIGNAL>
<PLACE type="WATER" id=2 form="NAM"> نهرالفرات </PLACE>
<PATH id=3 frame="VIEWER" signals="1"/>
<LINK id=4 source=3 target=2 linkType="EC">

나는 [유프라테스강][을] [따라] 여행했다.
I-Top Euphrates River-Acc along-ending travel-Past-ending
*along -> [을] [따라]
<SIGNAL id=1>을 </SIGNAL>
<SIGNAL id=2>따라</SIGNAL>
<PLACE type="WATER" id=3 form="NAM">유프라테스강</PLACE>
<PATH id=4 frame="VIEWER" signals="1"/>
<LINK id=5 source=3 target=2 linkType="EC"/>

□□ *[□□□□□□□□□□ □□] [□□□□] □□□□□□ □□□□□.*
<SIGNAL id=1> □□□□ </SIGNAL>
<PLACE type=WATER id=2 form=NAM> □□□□□□□□□□ □□ </PLACE>
<PATH id=3 frame=VIEWER signals="1"/>
<LINK id=4 source=3 target=2 linkType=EC>

我[沿着][幼发拉底河]旅行。
<SIGNAL id=1>*沿着*</SIGNAL>
<PLACE type="BODYOFWATER" id=2 form="NAM"> *幼发拉底河*</PLACE>
<PATH id=3 frame="VIEWER" signals="1"/>
<LINK id=4 source=3 target=2 linkType="EC">

17 Mapping to ACE

Mapping to ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 5.6.6 2006.08.01

In comparison with ACE, SpatialML attempts to use a classification scheme that's closer to information represented in gazetteers, thereby making the grounding of spatial locations in terms of geo-coordinates easier. SpatialML also doesn't concern itself with referential subtleties like metonymy; the latter has proven to be difficult for humans to annotate. Finally, SpatialML addresses relative locations involving distances and topological relations that ACE ignores. ACE 'GPE', 'Location', and 'Facility' Entity types are representable in SpatialML, as are ACE 'Near' Relations. Table 9 shows some example mappings for ACE entities, whereas Table 10 shows example mappings for ACE relations.

SpatialML, unlike ACE, is a 'flat' annotation scheme; Instead of grouping mentions into classes (called "entities" in ACE), SpatialML simply annotates mentions of places. Any mentions of ACE entities where the latter are of TYPE=GPE or TYPE=Location, or Facilities where SUBTYPE=Airports or SUBTYPE=Building-or-Grounds are candidate PLACE mentions, provided the ACE mentions have ROLE=GPE or ROLE=LOC and have ACE mention TYPE=NAM (i.e., proper names) or TYPE=NOM (nominals) are valid SpatialML PLACEs. Prenominal modifiers as in *the [US] population* are also considered PLACEs. Pronominal references such as *they*, *there*, *whose*, etc. are NOT considered PLACEs.

Text (SpatialML extents)	**SpatialML**	**ACE**
The continent of [Australia]	PLACE type="CONTINENT" continent="AU"	GPE type="CONTINENT"
the [Roman] emperor Constantine	PLACE type="PPLC" country="IT"	GPE type=Nation
[New York] Governor	PLACE type="CIVIL" state="NY" country="US"	GPE type="STATE"-or-Province
[Palm Beach] counties	PLACE type="CIVIL" state="FL" country="US"	GPE type=County-or-District
ABC news. [Washington].	PLACE type="PPLC" country="US"	GPE type=Population-Center
the [Middle East]	PLACE type="RGN"	GPE type=GPE-Cluster
[Palestine]	PLACE type="COUNTRY" country="PS"	GPE type=Special
met in [France]	PLACE type="COUNTRY" country="FR"	GPE.LOC
[Iraq] agreed to give	PLACE type="COUNTRY" country="IQ"	GPE.ORG

The rest of [America] voted	PLACE type="COUNTRY" country="US"	GPE.PER
pro-[Iraq] rally	PLACE type="COUNTRY" country="IQ"	GPE.GPE
the southern [United States]	PLACE type="RGN" mod="S" country="US"	Location
the center of the [city]	PLACE type="PPL" mod="C" ctv="CITY"	Location
[Capitol Hill]	PLACE type="PPL" state="DC" country="US"	Location type=Address
borders shared by [Turkey], [Azerbaijan], and [Georgia].	Three tags, with *Turkey, Azerbaijan,* and *Georgia* each annotated as type="COUNTRY"	Location type=Boundary
look directly at the [sun]	PLACE	Location type=Celestial
the [Missouri River]	PLACE type="WATER"	Location type=Water-Body
the southern [Caucasus]	PLACE type="RGN" mod="S"	Location type=Land-Region-natural
southern [Africa]	PLACE type="RGN" mod="S" continent=AF	Location type=Region-International
southern [Germany]	PLACE type="RGN" mod="S" country="DE"	Location type=Region-General
[La Guardia Airport]	PLACE type="FAC"	Facility type=Airport
[Disneyland]	PLACE type="FAC"	Facility type=Building-or-Grounds

Table 9: Mapping to ACE Entities

Mapping to ACE (Automatic Content Extraction) English Annotation Guidelines for Relations, Version 5.8.3 – 2005.07.01

ACE Relations of TYPE=PART-WHOLE.GEO or TYPE=PHYSICAL.NEAR are valid SpatialML Links. Our extent rules are different from ACE, which has generally longer and embedded tags as shown in Table 10.

Text (SpatialML extents)	**SpatialML**	**ACE**

[Moscow], [Russia]	PLACE type="PPLC" country="RU" id=1 PLACE type="COUNTRY" country="RU" id=2 LINK source=1 target=2 linkType="IN"	Relation: Part-Whole.GEO GPE Arg1: [Moscow, Russia] GPE Arg2: [Russia]
the top of the [mountain]	PLACE type="MTN" mod="T"	Relation: Part-Whole.GEO Location Arg1: [the top of the mountain] Location Arg2: [the mountain]
a [town] some [50 miles] [south] of [Salzburg] in the central [Austrian] [Alps]	*a [town] some [50 miles] [south] of [Salzburg] in the central [Austrian] [Alps]* <PLACE type="PPL" id=1 form="NOM" ctv="CITY">town</PLACE> <SIGNAL id=2>50 miles</SIGNAL> <SIGNAL id=3>south</SIGNAL> <PLACE id=4 type="PPLA" country="AU" form="NAM">Salzburg</PLACE> <PLACE id =5 type="COUNTRY" country="AT" mod="C">Austrian</PLACE> <PLACE id =6 type="MTS">Alps</PLACE> <PATH id=7 distance="50:mi" direction="S" source= 4 destination=1 signals="2 3"/> <LINK id=8 source=1 target=6 linkType="IN"/>	Relation: Physical.Near GPE Arg1: [a town some 50 miles south of Salzburg in the central Austrian Alps] GPE Arg2: [Salzburg]
the [Thai] border	PLACE type="COUNTRY" country="TH" mod="BR"	Relation: Part-Whole.GEO Location Arg1: [the Thai border] GPE Arg2: [Thai]
a military [base] in [Germany]	PLACE type="FAC" id=1 PLACE type="COUNTRY"	Relation: Part-Whole.GEO FAC Arg1: [a military base in

Note: The automatic conversion rules generate ACE extents (including embedded tags), rather than SpatialML extents. Further, the automatic conversion rules will over-generate in certain cases, e.g., "the town of X" will get marked as "the [town] of [X]". Still, they are far preferable to starting from scratch.

	country="GM" id=2 LINK source=1 target=2 linkType="IN"	Germany] GPE Arg2: [Germany]
[St. Vartan's Cathedral], on [Second Avenue]	PLACE type="FAC" id=1 PLACE type="ROAD" id=2 LINK source=1 target=2 linkType="IN"	Relation: Part-Whole.GEO FAC Arg1: [St. Vartan's Cathedral, on Second Avenue] FAC Arg2: [Second Avenue]
the [lobby] of the [hotel]	PLACE type="FAC" id=1 PLACE type="FAC" id=2 LINK source=1 target=2 linkType="IN"	Relation: Part-Whole.GEO FAC Arg1: [the lobby of the hotel] FAC Arg2: [the hotel]
the basketball [arena] of [Michigan State University]	PLACE type="FAC" id=1 PLACE type="FAC" id=2 LINK source=1 target=2 linkType="IN"	Relation: Part-Whole.GEO FAC Arg1: [the basketball arena of Michigan State University] FAC Arg2: [Michigan State University]

Table 10: Mapping to ACE Relations

18 Auto-Conversion of ACE data to SpatialML

A script has been developed to automatically convert ACE entity mentions and relations to possibly underspecified SpatialML PLACEs and LINKs. Tables 11 and 12 provide guidelines for mapping from SpatialML to ACE entities and relations respectively.

ACE Task	ACE Type	ACE Subtype	SpatialML convert
Entity	GPE		Place
	GPE	Continent	PLACE type="CONTINENT" continent= /string/
		Nation	PLACE

			type="COUNTRY" country=/string/
		State-or-Province	PLACE type="CIVIL"
		County-or-District	PLACE type="CIVIL"
		Population-Center	PLACE type="PPLC"
		GPE-Cluster	PLACE type="RGN"
		Special	PLACE type="COUNTRY" country= /string/
	Location		PLACE
		Celestial	PLACE type= "CELESTIAL"
		Water-Body	PLACE type= "WATER"
		Land-Region-natural	PLACE type="RGN"
		Region-International	PLACE type="RGN"
		Region-General	PLACE type="RGN"
	Facility	Airport	PLACE type="FAC"
		Building-or-Grounds	PLACE type="FAC"

Table 11: Rules for Automatically Mapping ACE Entities to SpatialML

ACE Task	**ACE Type**	**ACE Subtype**	**SpatialML convert**
Relation	PART-WHOLE	Geographical	LINK source=convert.id(/Role.Arg-1/) target=convert.id(/Role.Arg-2/) linkType="IN"
	Physical	Near	LINK source=convert.id(/Role.Arg-1/) target=convert.id(/Role.Arg-2/) linkType="NR"

Table 12: Rules for Automatically Mapping ACE Relations to SpatialML

19 Mapping to Toponym Resolution Markup Language (TRML)

Here is an example of TRML, from Leidner (2006):

```
<toponym term="BRUSSELS">
<candidates>
<cand id="c1" src="NGA" lat="-23.3833333" long="29.15"
humanPath="Brussels &gt; (SF04) &gt; South Africa" />
<cand id="c2" src="NGA" lat="-24.25" long="30.95"
humanPath="Brussels &gt; (SF04) &gt; South Africa" />
<cand id="c3" src="NGA" lat="-24.6833333" long="26.6833333"
humanPath="Brussels &gt; (SF04) &gt; South Africa" />
<cand id="c6" src="NGA" lat="50.8333333" long="4.3333333"
selected="yes"
humanPath="Brussels &gt; (BE02) &gt; Belgium" />
<cand id="c7" src="USGS_PP" lat="38.94944" long="-90.58861"
humanPath="Brussels &gt; Calhoun &gt; IL &gt; US &gt;
North America" />
</candidates>
</toponym>
```

In contrast to this approach, rather than having a list of candidate gazetteer references, we commit to a single one. If the place is ambiguous given the document as context, we do not list all gazetteer entries. However, within a tag, SpatialML optionally records latitude and longitude, where available, via a gazref as well as container information (corresponding to humanPath in TRML).

20 Mapping to GML

Most of the places represented in SpatialML can be represented in much richer detail in the OGC's GML, which is a soon-to-be ISO XML standard (ISO 19136) for marking up structured geographical data on the Web. (This can also support geographical calculations, display, etc.) Geo-coordinates for a given place, for example, can vary greatly, depending on what reference coordinate system and underlying geometric model of the earth (called a "geodetic" model) is being used. Further, even latitudes and longitudes may be provided in decimal units, or in degrees, minutes, and seconds. The precision may vary greatly when comparing across representations.

Fortunately, GML is highly expressive. For example, a geo-coordinate may be described as follows:

```
<gml:Point gml:name="Macy's" gml:id="3" srsName="urn:ogc:def:crs:EPSG:6.6:4326">
   <gml:coordinates>40.45 - 73.59</gml:coordinates>
</gml:Point>
```

This GML tag for Macy's says that the reference coordinate system is CRS 4326 (which happens to be the geodetic model WGS-84). It presents the coordinates in the format latitude followed by longitude (in this case in decimal degrees), with southern latitudes and western longitudes being expressed by negative signs. A richer tag might provide height and internal structure for Macy's as well.

As mentioned earlier, points are abstractions. Places construed as points can be represented, instead of by a geo-coordinate alone, as a circle centered on the geo-coordinate and a *radius of uncertainty* around that geo-coordinate. The following example shows a representation of Manhattan as a circle centered at Macy's and with a radius of 5000 meters.

```
  <gml:CircleByCenterPoint  gml:name="Manhattan">
        <gml:Point gml:name= "Manhattan" gml:id="3"
                srsName="urn:ogc:def:crs:EPSG:6.6:4326">
                 <gml:coordinates>40.45 - 73.59</gml:coordinates>
        </gml:Point>
   <gml:radius uom="urn:EPSG:uom:9001">5000</gml:radius>
</gml:CircleByCenterPoint>
```

One way of aligning a SpatialML tag with a GML representation is to wrap both in an XML based layer that has a tag that explicitly maps gml:id to SpatialML:id.

Thus, we might equate a PLACE tag for “5 miles east of Fengshan” with a particular GML tag corresponding to a coordinate with a particular area of uncertainty.

found in a [building] [5 miles] [east] of [Fengshan]

```
<PLACE type=“FAC” id=1 form=“NOM”>building</PLACE>

<SIGNAL id=2>5 miles</SIGNAL>

<SIGNAL  id=3>east</SIGNAL>

<PLACE type=“PPL” id=4 country=“TW” form=“NAM”  latLong=”2237N
12021E”>Fengshan</PLACE>

<PATH id=5 distance=“5:mi” direction=“E” source= 4 destination=1 signals=“2 3”/>


<gml:Point gml:id="3"  srsName="urn:ogc:def:crs:EPSG:6.6:4326">
   <gml:coordinates>22.66 120.41</gml:coordinates>
</gml:Point>
```

The wrapping layer will then equate SpatialML:id=1 with gml:id=3. This mapping may be generalized to PLACES of particular types. More commonly, however, there will be a transformation from one to the other that might be more complex.

Likewise, directions in SpatialML can be mapped to particular direction vectors with associated angles from a geo-coordinate in GML.

21 Mapping to KML

Keyhole Markup Language (KML) is the formatting language used by Google Earth to mark up geographical content on the Web for display using the Google Earth geographical browser. We illustrate a mapping using the same example as in the case of GML.

found in a [building] [5 miles] [east] of [Fengshan]

```
<PLACE type=FAC id=1 form="NOM">building</PLACE>
<SIGNAL id=2>5 miles</SIGNAL>
<SIGNAL  id=3>east</SIGNAL>
<PLACE type="PPL" id=4 country=TW form="NAM"  latLong="2262N
12034E">Fengshan</PLACE>
<PATH id=5 distance="5:mi" direction="E" source= 4 destination=1 signals="2 3"/>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.1">
<Folder>
<Placemark>
 <name>Fengshan</name>
 <description>Fengshan</description>
  <Point>
   <coordinates>120.35, 22.62</coordinates>
  </Point>
 </Placemark>
<Placemark>
 <name>building001</name>
 <description>building 5 miles east of Fengshan</description>
  <Point>
   <coordinates>120.42, 22.66</coordinates>
  </Point>
 </Placemark>
</Folder>
</kml>
```

Google Earth provides a rich set of display capabilities that can be scripted in KML. Thus*, a building 5 miles east of Fengshan* might be represented in KML by a point represented with an icon for a settlement, a line between that point and another point represented as a building, etc.

22 Towards SpatialML Lite

SpatialML will in all likelihood expand over subsequent versions, especially in covering other PLACE type and mod values. However, the DTD for SpatialML leaves every attribute of a PLACE tag except the tag id optional. This allows applications to decide which tags to use, and what attributes are needed. For example, a given application may choose only to include PLACE tags with latLong or gazref attributes. The specification of a lighter annotation scheme along these lines can be determined based on the needs of multiple applications.

23 SpatialML DTD

```
<!ELEMENT SpatialML ( #PCDATA | PLACE | PATH | LINK | SIGNAL )* >
<!ATTLIST SpatialML xsi:noNamespaceSchemaLocation CDATA #IMPLIED >
<!ATTLIST SpatialML xmlns:xsi CDATA #IMPLIED >
<!ATTLIST SpatialML comment CDATA #IMPLIED >

<!ELEMENT PLACE ( #PCDATA ) >
<!ATTLIST PLACE id ID #REQUIRED >
<!ATTLIST PLACE gazref CDATA #IMPLIED >
<!ATTLIST PLACE comment CDATA #IMPLIED >
<!ATTLIST PLACE type ( WATER | CELESTIAL | CIVIL | CONTINENT | COUNTRY | FAC | GRID |
                LATLONG | MTN | MTS | PPL | PPLA | PPLC | POSTALCODE | POSTBOX | RGN |
                ROAD | UTM | VEHICLE ) #IMPLIED >
<!ATTLIST PLACE mod ( B | BR | C | E | ENE | ESE | L | N | NE | NEAR | NNE | R | S | SE | SSE |
               SSW | SW | T | W | WSW ) #IMPLIED >
<!ATTLIST PLACE continent ( AF | AN | AI | AU | EU | GO | LA | NA | PA | SA ) #IMPLIED >
<!-- Country codes are ISO-3166-1 two-letters. -->
<!ATTLIST PLACE country ( AD | AE | AF | AG | AI | AL | AM | AN | AO | AQ | AR | AS | AT | AU |
                  AW | AX | AZ | BA | BB | BD | BE | BF | BG | BH | BI | BJ | BM | BN |
                  BO | BR | BS | BT | BV | BW | BY | BZ | CA | CC | CD | CF | CG | CH |
                  CI | CK | CL | CM | CN | CO | CR | CU | CV | CX | CY | CZ | DE | DJ |
                  DK | DM | DO | DZ | EC | EE | EG | EH | ER | ES | ET | FI | FJ | FK |
```

```
                FM | FO | FR | GA | GB | GD | GE | GF | GG | GH | GI | GL | GM | GN |
                GP | GQ | GR | GS | GT | GU | GW | GY | HK | HM | HN | HR | HT | HU |
                ID | IE | IL | IM | IN | IO | IQ | IR | IS | IT | JE | JM | JO | JP |
                KE | KG | KH | KI | KM | KN | KP | KR | KW | KY | KZ | LA | LB | LC |
                LI | LK | LR | LS | LT | LU | LV | LY | MA | MC | MD | ME | MG | MH |
                MK | ML | MM | MN | MO | MP | MQ | MR | MS | MT | MU | MV | MW | MX |
                MY | MZ | NA | NC | NE | NF | NG | NI | NL | NO | NP | NR | NU | NZ |
                OM | PA | PE | PF | PG | PH | PK | PL | PM | PN | PR | PS | PT | PW |
                PY | QA | RE | RO | RS | RU | RW | SA | SB | SC | SD | SE | SG | SH |
                SI | SJ | SK | SL | SM | SN | SO | SR | ST | SV | SY | SZ | TC | TD |
                TF | TG | TH | TJ | TK | TL | TM | TN | TO | TR | TT | TV | TW | TZ |
                UA | UG | UM | US | UY | UZ | VA | VC | VE | VG | VI | VN | VU | WF |
                WS | YE | YT | ZA | ZM | ZW | OTHER ) #IMPLIED >
<!ATTLIST PLACE form ( NAM | NOM ) #IMPLIED >
<!ATTLIST PLACE county NMTOKEN #IMPLIED >
<!-- State codes are ISO-3166-2  codes (excluding the country code and hyphen). -->
<!ATTLIST PLACE state NMTOKEN #IMPLIED >
<!ATTLIST PLACE latLong CDATA #IMPLIED >
<!ATTLIST PLACE nonLocUse ( true | false ) #IMPLIED >
<!ATTLIST PLACE description CDATA #IMPLIED >
<!ATTLIST PLACE CTV ( CITY | TOWN | VILLAGE ) #IMPLIED >

<!ELEMENT PATH EMPTY >
<!ATTLIST PATH id ID #REQUIRED >
<!ATTLIST PATH comment CDATA #IMPLIED >
<!ATTLIST PATH source IDREF #REQUIRED >
<!ATTLIST PATH destination IDREF #REQUIRED >
<!ATTLIST PATH signals IDREFS #REQUIRED >
<!ATTLIST PATH frame ( VIEWER | INTRINSIC | EXTRINSIC ) #IMPLIED >
<!ATTLIST PATH direction ( B | A | BL | E | F | N | S | W | NNE | NE | ENE | ESE | SE | SSE |
                SSW | SW | WSW | WNW | NW ) #IMPLIED >
<!ATTLIST PATH distance CDATA #IMPLIED >
```

```
<!ELEMENT LINK EMPTY >
<!ATTLIST LINK id ID #REQUIRED >
<!ATTLIST LINK comment CDATA #IMPLIED >
<!ATTLIST LINK source IDREF #REQUIRED >
<!ATTLIST LINK target IDREF #REQUIRED >
<!ATTLIST LINK linkType ( IN | EC | EQ | DC | PO| NR ) #IMPLIED >

<!ELEMENT SIGNAL ( #PCDATA ) >
<!ATTLIST SIGNAL id ID #REQUIRED >
<!ATTLIST SIGNAL comment CDATA #IMPLIED >
```

24 Future Work

- Mapping to spatial upper model ontologies, such as found in SUMO
- Other kinds of MODs.
- Standardizing States.
- More extensive topological relations.
- Sets of Locations e.g., *all cities that have a population more than five million.*
- Representing uncertainty

References

Cohn, A. G., Bennett, B., Gooday, J., Gotts, N. M. 1997. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. GeoInformatica, 1, 275–316, 1997.

Garbin, Eric and Inderjeet Mani. 2005. Disambiguating Toponyms in News. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 363–370. Association for Computational Linguistics, Vancouver, British Columbia, Canada.

Leidner, Jochen L. 2006. Toponym Resolution: A First Large-Scale Comparative Evaluation. Research Report EDI-INF-RR-0839 (July 2006).

Mardis, Scott and John Burger. 2005. Design for an Integrated Gazetteer Database

Technical Description and User Guide for a Gazetteer to Support Natural Language Processing Applications. MITRE TECHNICAL REPORT, MTR 05B0000085, November 2005

Randell, D. A., Z. Cui, and A. G. Cohn. 1992. A Spatial Logic Based on Regions and Connection, Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, pp. 165–176, 1992.

Schilder, Frank Versley, Y., & Habel, C. 2004. Extracting Spatial Information: Grounding, Classifying and Linking Spatial Expressions. In the Workshop on Geographic Information Retrieval at the 27th ACM SIGIR conference, Sheffield, England, UK.