SpatioTemporal MITRE-Sponsored Research

SpatialML:

Annotation Scheme for Marking Spatial Expressions in Natural Language

June 28, 2010

Version 3.0.1

Contact: cdoran@mitre.org

Research funded by the MITRE Innovation Program.
Approved for Public Release; Distribution Unlimited. Case 07-0614.

Annotation Scheme for Marking....1
Spatial Expressions 1
in Natural Language....1
June 28, 2010....1
Version 3.0.1....1
Acknowledgements....3
1Introduction....4
2Changes since version 2.0....4
3Building on Prior Work....4
4Extent Rules (English-specific)....4
5Toponyms....5
5.1Mapping Continents, Countries, and Country Capitals....5
5.2Mapping via Gazetteer Unique Identifiers....12
5.3Mapping via Geo-Coordinates....13
5.4UnMappable Places....13
6Ambiguity in Mapping....13
6.1Ambiguity in Text....14
6.2Genuine Ambiguity in the Gazetteer....14
6.3Multiple Gazetteer Entries for the Same Place....14
6.4When the Gazetteer is too Fine-Grained Compared to Text....14
7Mapping Restrictions via the MOD attribute....15
8Using the Type Feature....17
9Annotating Geo-Coordinates found in text....19
10Annotating Addresses....19
11Marking Exceptional Information....19
12Annotating Relative Locations via Spatial Relations....19
12.1LINKs vs. RLINKs....19
12.2 LinkTypes....22
13Non-consuming PLACE tags....25
14Disambiguation Guidelines....25
15States....26
16Inventory of SpatialML Tags....26
17Multilingual Examples....30
18Mapping to ACE 39
19Auto-Conversion of ACE data to SpatialML 41
20Mapping to Toponym Resolution Markup Language (TRML)....42
21Mapping to GML 42
22Mapping to KML 43
23Mapping to GUM....44
24Towards SpatialML Lite....45
25SpatialML Version....45
26SpatialML DTD (Version 3.0)....45
27Future Work....48

References.....49

Research funded by the MITRE Innovation Program.
Approved for Public Release; Distribution Unlimited. Case 07-0614.

Acknowledgements

SpatialML 3.1 is the latest 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 3.1 or prior versions:

- John Bateman (University of Bremen)
- Cheryl Clark (MITRE)
- Christy Doran (MITRE)
- Jade Goldstein-Stewart (Department of Defense)
- Dan Parvaz (MITRE)
- Dave Harris (MITRE)
- Dulip Herath (University of Colombo)
- Qian Hu (MITRE)
- Janet Hitzeman (MITRE)
- Seok Bae Jang (Georgetown University)
- Inderjeet Mani (MITRE)
- Karine Megerdoomian (MITRE)
- Ross Purves (University of Zurich)
- James Pustejovsky (Brandeis University)
- Justin Richer (MITRE)
- Seamus Clancy (MITRE)
- Scott Mardis (MITRE)

This version will be posted at:
http://sourceforge.net/projects/spatialml

We expect that subsequent releases will continue to 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. Throughout the document, example text is called out in a box such as:

This is a piece of example text

[along with] any annotations [it may have].

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 RLINK 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 25.

2 Changes since version 2.0

- PATH is now RLINK ("relative link").
- The word "in" is now a SIGNAL when it motivates a LINK.[1]
- The new TYPE attribute to SIGNALs has possible values of "DIRECTION" and "DISTANCE".
- EC, which stood for "extended connection," now stands for "external connection", to be consistent with RCC8 terminology.

1 Optionally, the comma in "Boston, MA" can also be tagged as a SIGNAL. We have not done so in this corpus.

- The CTV (city/town/village) feature has been eliminated.
- NEAR is no longer a linkType. The word “near” is marked as a SIGNAL, the token NEAR also is the value of a MOD.
- In phrases such as “city of Baton Rouge”, “city” is now tagged as a separate PLACE, and there is a LINK between “city” with “Baton Rouge” with a linkType of “EQ”.
- There is now a version attribute to the root SpatialML tag to capture which version of the guidelines files have been annotated with.
- MOD and DIRECTION features have all been expanded to full words (cf. Tables 3 and 5); additional compass points added (ENE, etc.), as well as additional MOD values TOP, CENTRAL, LEFT and RIGHT.
- Codes for MODs, Directions, and Distances have been mapped to the Bremen Generalized Upper Model (GUM) Ontology[2] classes.
- There is now a boolean PREDICATIVE feature on PLACE tags.

3 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[3] 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[4] and the 2005 Automatic Content Extraction (ACE) guidelines.[5]
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.[6] We also leverage the integrated

2http://www.ontospace.uni-bremen.de/linguisticOntology.html

3http://timex2.mitre.org

4http://www.timeml.org

5http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html

6http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/top.htm.

gazetteer database (IGDB) of (Mardis and Burger 2005). Last but not least, this annotation scheme can be related to the Geography Markup Language (GML)[7] defined by the Open Geospatial Consortium (OGC), as well as Google Earth's Keyhole Markup Language (KML)[8] 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 RCC 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.

4 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 phrases such as "here".
- PLACEs can be in the form of proper names ("New York") or nominals ("town"), which are marked with the "form" attribute as NAM or NOM, respectively.
- 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."[9]
- Expressions such as "city" in "the city of Baton Rouge" are tagged, and are then LINKed to their city (in this case "Baton Rouge") via an EQ link; "city" is a NOM while "Baton Rouge" is a NAM.

7http://www.opengis.net/gml/

8http://earth.google.com/kml/

9 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.
- Periods are included in the tag extent in the case of abbreviations, and the *'s* of possessives are also included in the tag extent.
- 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 RLINK) to express relationships between PLACEs.
- In the corpus we are releasing, we also tag facilities (PLACEs with type FAC) as defined by the ACE guidelines and provide LINKs between FACs and their locations.

5 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).[10] 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.[11] The consolidation is done in the IGDB gazetteer (Mardis and Burger 2005) developed at MITRE for the Disruptive Technologies Office.

5.1 Mapping Continents, Countries, and Country Capitals

How redundant your gazetteer is will determine which entries require a gazref and/or latlong feature. IGDB is highly redundant, so when using it, we always use a gazref when we can find one.The values COUNTRY, CONTINENT, and PPLC for the type feature are sufficient to disambiguate the corresponding PLACEs for many gazetteers.

10 http://nhd.usgs.gov/gnis.html

11 http://gnswww.nga.mil/geonames/GNS/index.jsp

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>

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.

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).[12] In our annotation, we have chosen to go with the ISO 3166-2 option, but this is an arbitrary choice made for consistency. Similarly, when Australia is mentioned, we have chosen to annotate it as a country rather than a continent, solely for consistency.

12 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[13]	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

13 In 3166-2, the ISO standard for provinces/states, Hong Kong is listed as CN-91. We must expect some inconsistencies as the standards are updated, and we must expect that the standards will have to be updated as country names and borders change.

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.)

5.2 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 in the form of a UNI. For example, IGDB references are uniquely identified by an integer; as such, the IGDB reference #912321 would be listed as “IGDB:912321” in a gazref attribute.

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>
<PLACE is=3 type=“PPLA” country=“IN” form=“NOM” gazref=“IGDB:17896959”
>city</PLACE>
<SIGNAL id=4>in</SIGNAL>
<SIGNAL id=5>of</SIGNAL>
<LINK source=2 target=1 signals=“4” linkType=“IN”>
<LINK source=3 target=1 signals=“5” linkType=“EQ”>

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

5.3 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.

In 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.

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="US-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.

5.4 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>

6 Ambiguity in Mapping

6.1 Ambiguity in Text

It may often be the case that the document 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 document is genuinely ambiguous, we tag the place without any gazetteer reference or geo-coordinate.

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

6.2 Genuine Ambiguity in the 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="US-IL" country="US" form="NAM">Rochester</PLACE>
<PLACE id=2 state="US-IL" country="US" type="STATE" form="NAM"
gazref="IGDB:30125575">Illinois</PLACE>
<SIGNAL id=3>in</SIGNAL>
<LINK source=1 target=2 signals="3" 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.

6.3 Multiple Gazetteer Entries for the Same Place

When there is more than one correct 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 there are still multiple choices, 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.[14] One can also extract the latlong from that entry via script as a post-process to eliminate one source of spurious human error.
In any specific annotation task, you may want to agree by convention on certain features for highly ambiguous cases, e.g the gazetteer entry for United States, whether Russia is in Asia or Europe, which continent the Caribbean countries are in, or which of the many entries to use for Washington, DC.

6.4 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].

14 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, e.g., "Indiana,%". The result will give latlongs for PLACEs such as "The Commonwealth of Massachusetts" and "The Kingdom of The Netherlands."

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.

If we can't resolve the ambiguity in the gazetteer, we leave out the gazref and geo-coordinate

```
<PLACE id=1 state="US-IL" country="US" form="NOM">town</PLACE>
<PLACE id=2 state="US-IL" country="US" form="NAM">Rochester</PLACE>
<PLACE id=3 state="US-IL" country="US" type="RGN"
        form="NAM">Sangamon County</PLACE>
<PLACE id=4 state="US-IL" country="US" type="STATE" form="NAM"
        gazref="IGDB:30125575">Illinois</PLACE>
<SIGNAL id=5>of</SIGNAL>
<SIGNAL id=6>in</SIGNAL>
<LINK id=7 source=1 target=2 signals="5" linkType="EQ"/>
<LINK id=8 source=2 target=3 signals="6" linkType="IN"/>
<LINK id=9 source=3 target=4 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.

If the gazetteer supports equivalence class filtering, pick the first gazref in the equivalence class.

7 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>
```

Note that mods never have a tagged extent, unless the mod is part of the PLACE name, as in the example below.

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, which include a sixteen-point compass rose. 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.

BOTTOM	the bottom of the [well]
BORDER	[Burmese] border
NEAR	near [Harvard]
TOP	the top of the [mountain]
CENTRAL	central Asia
LEFT	the left side of the house
RIGHT	the right side of the car
N	northern India
NNE	north-northeast
NE	northeast
ENE	east-northeast
E	eastern India
ESE	east-southeast
SE	southeast
SSE	south-southeast
S	southern India
SSW	south-southwest
SW	southwest
WSW	West-southwest
W	west Tikrit
WNW	West-northwest
NW	northwest
NNW	north-northwest

Table 3: MOD Codes

8 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.
As mentioned above, if the types CONTINENT, COUNTRY, STATE and LATLONG are used, the corresponding slots continent, country, state and latlong must be filled *only* if they are not specified by the gazref entry.

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. provinces, certain instances of towns and cities, the European Union,
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, post codes, 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, e.g. desert, suburbs, beach
ROAD	Street, road, highway, etc.
STATE	A first-order administrative division within a country, e.g., state, gubernia, territory, etc. States are entities which have ISO 3166-2 codes. See Table 7.
UTM	A Universal Transverse Mercator (UTM) format indication of the location
VEHICLE	Car, truck, train, etc.

Table 4: TYPE Codes

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, an 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"/>
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

Each portion of an address should be marked separately.
[100 James Drive, SE], [Vienna], [Virginia] [22180]
<PLACE type="ROAD">100 James Drive, SE</PLACE>
<PLACE type="PPL" state="US-VA">Vienna</PLACE>
<PLACE type="STATE" state="US-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*.
A feature since version 2.1 of the guidelines is the predicative flag. This is set to "true" for cases in which the PLACE phrase is adjectival (including overt possessive, compound nouns and bare adjectival forms) and therefore predicates a property onto an object. Examples of such cases are "Iraqi soldier", "US President", "Europeans" and "American athletes." We want to capture the fact that there is a relationship between "Iraq" and the "soldier," namely that the soldier fights for Iraq even though he is not necessarily in Iraq. This feature is used merely to indicate that a relationship exists, but does not specify what that relationship is. We have chosen this approach to avoid asking the annotator to untangle the semantic relationships within complex nominals, such as "Norwegian pleasure cruise." In this example, we don't want the annotator to have to specify whether the pleasure is Norwegian in nature.

12 Annotating Relative Locations via Spatial Relations

12.1 LINKs vs. RLINKs

We use an RLINK 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 type="DIRECTION">northwest</SIGNAL>
<RLINK direction="NW" source=2 target=1 signals="3" />

The RLINK 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 RLINK. 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 RLINK may be included in a signals attribute in the RLINK tag.
a [town] some [50 miles] [south] of [Salzburg] [in] the central [Austrian] [Alps]
<PLACE type="PPL" id=1 form="NOM">town</PLACE>
<SIGNAL id=2 type="DISTANCE">50 miles</SIGNAL>
<SIGNAL id=3 type="DIRECTION">south</SIGNAL>
<PLACE id=4 type="PPLA" country="AT" form="NAM">Salzburg</PLACE>
<PLACE id=5 type="COUNTRY" country="AT" mod="CENTER">Austrian</PLACE>
<PLACE id=6 type="MTS" >Alps</PLACE>
<RLINK id=7 distance=2 direction="S" source=4 target=1 signals="2 3"/>
<SIGNAL id=8>in</SIGNAL>
<LINK id=9 source=1 target=6 signals=8 linkType="IN"/>
<LINK id=10 source=6 target=5 linkType="IN"/>
Mark RLINKs only when they are described within one phrase, i.e., if parts of a RLINK are described in different sentences or in different parts of the same sentence, do not mark them.
For direction codes, refer to Table 5, which includes several relational directions in addition to a sixteen-point compass rose (such as found in PLACE MODs in Table 3).

Direction	**Example**
ABOVE	[above] the roof
BELOW	[below] the tree-line
FRONT	[in front of] the theater
BEHIND	[behind] the house
N	[north] of
NNE	north-northeast
NE	northeast
ENE	east-northeast
E	[east] of
ESE	east-southeast
SE	southeast
SSE	south-southeast
S	[south] of
SSW	south-southwest
SW	southwest
WSW	West-southwest

W	[west] of
WNW	West-northwest
NW	northwest
NNW	north-northwest

Table 5: Codes for Directions

12.2 LinkTypes

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) to indicate that “town” is in “Alps”. The set of linkTypes (Table 6) is derived in part from the Region Connection Calculus (Randell et al. 1992, Cohn et al. 1997). DC, EC, EQ, and PO are from the calculus version known as RCC8. IN is not RCC8, but collapses two RCC8 relations, TPP and NTPP (tangential proper part and non-tangential proper part, respectively)[15]. The reason for the collapsing is that it is often difficult for annotators to decide whether the “part” touches or doesn't touch the container's border. Finally, we don't include the remaining RCC8 inverse links TPPi and NTPPi from RCC8, since these can be represented in annotation by swapping arguments, and are in addition likely to confuse annotators.

15 IN is in fact the PP "proper part" relation in RCC5.

LinkType	Example
IN (tangential and non-tangential proper parts)	[Paris], [Texas]
EC (external connection)	the border between [Lebanon] and [Israel]
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>
<SIGNAL id=3>in</SIGNAL>
<LINK source=1 target=2 signals="3" linkType="IN"/>

this northern [Uganda] [town]

<PLACE type="PPL" country="UG" id=1>town</PLACE>
<PLACE type="COUNTRY" country="UG" mod="N" id=2>Uganda</PLACE>
<LINK source=1 target=2 linkType="IN"/>

The [US]-[Canadian] border

<PLACE type="COUNTRY" country="US" id=1>US</PLACE>
<PLACE type="COUNTRY" country="CA" id=2>Canadian</PLACE>
<LINK source=1 target=2 linkType="EC"/>

[Pacific] coast of [Australia]

<PLACE type="WATER" mod="BORDER" id=1>Pacific</PLACE>
<PLACE type="COUNTRY" country="AU" id=2>Australia</PLACE>
<LINK source=1 target=2 linkType="EC"/>

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

<PLACE type="RGN" mod="CENTER" id=1 form="NOM">district</PLACE>
<PLACE type="PPL" id=2 form="NOM">town</PLACE>
<PLACE type="PPL" id=3 form="NAM">Tirunelveli</PLACE>
<PLACE type="CIVIL" country="IN" id=4 form="NAM">Tamil Nadu</PLACE>
<PLACE type="COUNTRY" country="IN" mod="S" id=5 form="NAM">India</PLACE>
<SIGNAL id=6>of</SIGNAL>
<SIGNAL id=7>of</SIGNAL>
<SIGNAL id=8>in</SIGNAL>
<LINK source=1 target=2 signals="6" linkType="IN"/>
<LINK source=2 target=3 signals="7" linkType="EQ"/>
<LINK source=4 target=5 signals="8" 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>
<RLINK id=5 direction="BEHIND" source=1 target=3 frame="VIEWER" signals="2"/>
<SIGNAL id=6 >on</SIGNAL
<LINK id=7 source=3 target=4 signals="6" linkType="IN"/>

[towards] [Scammonden Water] [along] the [B6114]
<SIGNAL id=1>towards</SIGNAL>
<PLACE type="WATER" country="GB" id=2
form="NAM">Scammonden Water</PLACE>
<SIGNAL id=3>along</SIGNAL>
<PLACE type="ROAD" country="GB" id=4 form="NAM">B6114</PLACE>
<RLINK id=5 target=2 frame="VIEWER" signals="1 3"/>
<LINK id=6 source=5 target=4 linkType="EC">

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

13 Non-consuming PLACE tags

As of version 3.0, we have added an optional feature for non-consuming PLACE tags in order to handle cases such as the following:

We drove 50 miles east of Boston.

This example in itself is fine and can be treated as a typical RLINK from *[Boston]* in the direction east for the distance 50 miles. Suppose the second sentence is the following, however:

The next day, we drove [100 miles] [north].

We know the travelers were in two places, but those two places don't have PLACE marks. For purposes of tracking travel routes, etc., having these PLACE marks is desirable.

The annotation of these two sentences is:

We drove [50 miles] [east] of [Boston]. The next day, we drove [100 miles] [north].
<PLACE id=1 type="PPLA" country="US" state="US-MA">Boston</PLACE>
<SIGNAL id=2 type="DISTANCE">50 miles</SIGNAL>
<SIGNAL id=3 type="DIRECTION">east</SIGNAL>
<PLACE id=4/>
<RLINK id=5 source=1 target=4 distance=2 direction="E" signals= "2 3">
<SIGNAL id=6 type="DISTANCE">100 miles</SIGNAL>
<SIGNAL id=7 type="DIRECTION">north</SIGNAL>
<PLACE id=9/>
<RLINK id=10 source=4 target=8 distance=6 direction="N" signals= "6 7">

14 Disambiguation Guidelines

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.

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

  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.

15 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 in SpatialML are defined by ISO 3166-2 codes. (See http://en.wikipedia.org/wiki/ISO_3166-2 or http://www.iso.org/iso/country_codes/background_on_iso_3166/iso_3166-2.htm), When entering a state code, one can enter either the country code followed by a dash and then the state code (e.g., US-WA, or FR-91) or enter just the state code (e.g., WA or 91) and fill in the country code in the "country" attribute field.

State codes are ISO 3166-2 codes. For states not found in the definition for ISO 3166-2, a Wikipedia or Google search for the name of the state is often helpful.

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

16 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
	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
RLINK	source	tagid
	id	tagid
	target	tagid
	direction	tagid
	distance	tagid
	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
	type	DISTANCE or DIRECTION, or else this attribute is omitted
	comment	text field

Table 8: SpatialML Tags and Attributes

17 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. One issue to be determined is whether SIGNALs such as “of” and “in” are present in other languages, and what form they take, e.g., preposition, affix or morpheme.

1. I attended a pro-[American] rally.

<PLACE type=“COUNTRY” country=“US” form=“NAM”>American</PLACE>

Here is the corresponding Arabic.

حضرت مظاهرة مريدة- للوللا يات المتحدة

<PLACE type=“COUNTRY” country=“US” form=“NAM”> الولايات المتحدة</PLACE>

Turning to Korean:

□□ □□-[□□□□] □□□ □□□□□.

<PLACE type=“COUNTRY” country=“US” form=“NAM”>□□□□</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>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> اوغندأ</PLACE>

<LINK source=1 target=2 linkType=“IN”/>

□□ □ [□□□] □□ [□□]□ □□.
<PLACE type="COUNTRY" country="UG" mod="N" id=1>□□□</PLACE>
<PLACE type="PPL" country="UG" id=2>□□</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.
□□ □□□ □□□□□ [□□□□□□] [□□□□□□□]
□□□□□□.
<PLACE type="PPL" country="UG" id=1> □□□□□□□ </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>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 type="DIRECTION">northwest</SIGNAL>
<RLINK direction="NW" source=2 target=1 signals="3" />

أنا أسكن فى امريتسار شمال غرب العاصمة نيودلهى
<PLACE type="PPL" id=1 country="IN" form="NAM"> امريتسار</PLACE>
<PLACE type="PPLC" country="IN" id=2 form="NAM"> نيودلهى</PLACE>
<SIGNAL id=3 type="DIRECTION"> شمال غرب</SIGNAL>
<RLINK direction="NW" source=2 target=1 signals="3"/>

□□ □□ [□□□] [□□□]□ [□□□□]□ □□.
<PLACE type="PPLC" country="IN" id=1 form="NAM">□□□</PLACE>
<SIGNAL id=2 type="DIRECTION">□□□</SIGNAL>
<PLACE type="PPL" id=3 country="IN" form="NAM">□□□□</PLACE>
<RLINK direction="NW" source=1 target=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 type="DIRECTION"> □□□□□□ </SIGNAL>
<RLINK direction="NW" source=2 target=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 type="DIRECTION">西北部 </SIGNAL>
<RLINK direction="NW" source=2 target=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">town</PLACE>
<SIGNAL id=2 type="DISTANCE">50 miles</SIGNAL>
<SIGNAL id=3 type="DIRECTION">south</SIGNAL>
<PLACE id=4 type="PPLA" country="AT" form="NAM">Salzburg</PLACE>
<PLACE id =5 type="COUNTRY" country="AT" mod="CENTER">Austrian</PLACE>
<PLACE id =6 type="MTS" >Alps</PLACE>
<RLINK id=7 distance=2 direction="S" source=4 target=1 signals="2 3"/>
<SIGNAL id=8>in</SIGNAL>
<LINK id=9 source=1 target=6 signals="8" linkType="IN"/>
<LINK id=10 source=6 target=5 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 type="DISTANCE"> خمسين ميل</SIGNAL>
<SIGNAL id=3 type="DIRECTION"> جنوب/SIGNAL>
<PLACE id=4 type="PPLA" country="AT" form="NAM"> سالزبرج</PLACE>
<PLACE id=5 type="COUNTRY" country="AT" mod="CENTER"> النمسا<PLACE>
<PLACE id =6 type="MTS"> جبال الالب </PLACE>
<RLINK id=7 distance=2 direction="S" source= 4 target=1 signals="2 3"/>
<SIGNAL id=8> فى</SIGNAL>
<LINK id=9 source=1 target=6 signals="8" linkType="IN"/>
<LINK id=10 source=6 target=5 linkType="IN"/>

□□ [□□□□□□] [□□□□] □□□□ [□□□□□□] [□□]□□ [50 □□] □□□□ [□□]□ □□.
<PLACE id =1 type="COUNTRY" country="AT" mod="CENTER">□□□□□□</PLACE>
<PLACE id =2 type="MTS" >□□□□</PLACE>
<PLACE id=3 type="PPLA" country="AT" form="NAM">□□□□□□</PLACE>
<SIGNAL id=4 type="DIRECTION">□□</SIGNAL>
<SIGNAL id=5 type="DISTANCE">50 □□</SIGNAL>
<PLACE type="PPL" id=6 form="NOM">□□</PLACE>
<RLINK id=7 distance=5 direction="S" source= 3 target=6 signals="4 5"/>
<LINK id=8 source=6 target=2 linkType="IN"/>

<LINK id=9 source=2 target=1 linkType="IN"/>

□□ □□□
□□ [□□□□□
□□□□□□] [□□□□□□□□] □□□□□
□□□ □□□□□□ [□□□□□□□□□□□□□□□□□□□□] [□□□□□□□□ 50 □□] □□□ [□□□□□□□□] □□□□□□
[□□□□□]
□□□□□□.
<PLACE type="PPL" id=1 form="NOM"> □□□□□ </PLACE>
<SIGNAL id=2 type="DISTANCE"> □□□□□□□□ 50 □□ </SIGNAL>
<SIGNAL id=3 type="DIRECTION"> □□□□□□□□ </SIGNAL>
<PLACE id=4 type="PPLA" country="AT" form=NAM> □□□□□□□□□□□□□□□□□□□□
</PLACE>
<PLACE id =5 type="COUNTRY" country="AT" mod="CENTER"> □□□□□
□□□□□□ </PLACE>
<PLACE id =6 type="MTS"> □□□□□□□ </PLACE>
<RLINK id=7 distance=2 direction="S" source= 4 target=1 signals="2 3"/>
<LINK id=8 source=1 target=6 linkType="IN"/>
<LINK id=9 source=6 target=5 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="AT" mod="CENTER"> 奥地利 </PLACE>
<PLACE id =2 type="MTS" >阿尔卑斯</PLACE>
<PLACE id=3 type="PPLA" country="AT" form="NAM"> 萨尔茨堡 </PLACE>
<SIGNAL id=4 type="DIRECTION">以南</SIGNAL>
<SIGNAL id=5 type="DISTANCE">50 英哩</SIGNAL>
<PLACE type="PPL" id=6 form="NOM"> 镇子</PLACE>
<RLINK id=7 distance=5 direction="S" source=3 target=6 signals="4 5"/>
<LINK id=8 source=6 target=2 linkType="IN"/>
<LINK id=9 source=2 target=1 linkType="IN"/>

5. I met Laila in a [cafè] [in] [Rabat].
<PLACE id=1 form="NOM" type="FAC">cafe</PLACE>
<PLACE type="PPLC" id=2 form="NAM" country="MA">Rabat</PLACE>
<SIGNAL id=3>in</SIGNAL>
<LINK id=4 source=1 target=2 signals="3" linkType="IN"/>

ألتقيت بليلى في مقهي في الرباط
<PLACE id=1 form="NOM" type="FAC"> مقهي</PLACE>
<PLACE type="PPLC" id=2 form="NAM" country="MA"> الرباط</PLACE>
<SIGNAL id=3> في </SIGNAL>

<LINK id=4 source=1 target=2 signals="3" 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"/>

□□ [□□□□□□] [□□□□□□□] □□□□□ □□□□□□□.
<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>
<SIGNAL id=3 >个</SIGNAL>
<LINK id=4 source=2 target=1 signals="3" linkType="IN"/>

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

أنا أسكن في المدينة العراقية الحدودية الرئيسية قم
<PLACE type="COUNTRY" id=1 form="NAM" mod="BORDER" country="IQ"> المدينة العراقية</PLACE>
<PLACE type="PPL" id=2 form="NAM" country="IQ"> قم </PLACE>
<LINK id=4 source=2 target=1 linkType="IN"/>

□□ □□□ [□□] □ □□ □□□ [□□]□ □□.
<PLACE type="COUNTRY" id=1 form="NAM" mod="BORDER" country="IQ">□□□
</PLACE>
<PLACE type="PPL" id=2 form="NAM" country="IQ">□□</PLACE>
<LINK id=3 source=2 target=1 linkType="IN"/>

□□ □□
□□□□ [□□□□] □□□□□□□□ □□□□ □□ [□□□□□□□□□□]
□□□□□□.
<PLACE type="COUNTRY" id=1 form="NAM" mod="BORDER" country="IQ"> □□□□
</PLACE>

<PLACE type="PPL" id=2 form="NAM" country="IQ"> □□□□□□□□□□ </PLACE>
<LINK id=4 source=2 target=1 linkType="IN"/>

我住在[伊拉克]边界的重[镇][奎坶]。
<PLACE type="COUNTRY" id=1 form="NAM" mod="BORDER" country="IQ"> 伊拉克 </PLACE>
<PLACE type="PPL" id=2 form="NAM" country="IQ"> 奎坶</PLACE>
<LINK id=4 source=2 target=1 linkType="IN"/>

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 type="DISTANCE">200 miles</SIGNAL>
<SIGNAL id=3 type="DIRECTION">west</SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ">Baghdad</PLACE>
<RLINK id=5 distance=2 direction="W" source=4 target=1 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 type="DISTANCE"> مائتين ميلا </SIGNAL>
<SIGNAL id=3 type="DIRECTION">غرب</SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ"> بغداد</PLACE>
<RLINK id=5 distance=2 direction="W" source=4 target=1 signals="2 3"/>

□□ [□□□□] [□□]□□ □ 200 □□ □□□ [□□]□□ □□□□.
<PLACE type="PPL" id=1 form="NAM" country="IQ">□□</PLACE>
<SIGNAL id=2 type="DISTANCE">200 □□</SIGNAL>
<SIGNAL id=3 type="DIRECTION">□□</SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ">□□□□</PLACE>
<RLINK id=5 distance=2 direction="W" source=4 target=1 signals="2 3"/>

□□ [□□□□□□□□□□] [□□□□□□□□ 200 □□] □□□ [□□□□□□□□□] □□□□□□ [□□□□□□□□□□□] □□□□□□□□.
<PLACE type="PPL" id=1 form="NAM" country="IQ"> □□□□□□□□□□ </PLACE>
<SIGNAL id=2 type="DISTANCE"> □□□□□□□□ 200 □□ </SIGNAL>
<SIGNAL id=3 type="DIRECTION"> □□□□□□□□□ </SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ"> □□□□□□□□□□ </PLACE>
<RLINK id=5 distance=2 direction="W" source=4 target=1 signals="2 3"/>

我出生在离[巴格达][西面]大约[二百英哩]的[奎坶]。
<PLACE type="PPL" id=1 form="NAM" country="IQ"> 巴格达 </PLACE>
<SIGNAL id=2 type="DIRECTION "> 西面 </SIGNAL>

<SIGNAL id=3 type="DISTANCE" ">二百英里 </SIGNAL>
<PLACE type="PPLC" id=4 form="NAM" country="IQ"> 奎坶 </PLACE>
<RLINK id=5 distance=2 direction="W" source=4 target=1 signals="2 3"/>

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

أنا أسكن علي بعد ما يقارب من أثنين ميل من حدود المكسيك
<SIGNAL id=1 type="DISTANCE"> أثنين ميل </SIGNAL>
<PLACE type="COUNTRY" id=2 form="NAM" mod="BORDER"
country="MX"> المكسيك </PLACE>
<RLINK id=3 distance=1 source=2 signals="1"/>

□□ [□□□] □□□□ [2 □□] □□ □□.
<PLACE type="COUNTRY" id=1 form="NAM" mod="BORDER" country="MX">□□□
</PLACE>
<SIGNAL id=2 type="DISTANCE">2 □□</SIGNAL>
<RLINK id=3 distance=2 source=1 signals="2"/>

□□ [□□□□□□□□□□]
□□□□□□□□□□ □□□ [□□□□□□□ 2 □□] □□□□□
□□□□□□.
<SIGNAL id=1 type="DISTANCE">□□□□□□□ 2 □□ </SIGNAL>
<PLACE type="COUNTRY" id=2 form="NAM" mod="BORDER"
country="MX">□□□□□□□□□□ </PLACE>
<RLINK id=3 distance=1 source=2 signals="1"/>

我住在离[墨西哥]边境[两英哩]以内。
<PLACE type="COUNTRY" id=1 form="NAM" mod="BORDER"
country="MX"> 墨西哥 </PLACE>
<SIGNAL id=2 type="DISTANCE">两英里 </SIGNAL>
<RLINK id=3 distance=2 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>
<RLINK 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>
<RLINK id=3 frame="VIEWER" signals="1"/>
<LINK id=4 source=3 target=2 linkType="EC"/>

□□ [□□□□□□][□] [□□] □□□□.
<PLACE type="WATER" id=1 form="NAM">□□□□□□</PLACE>
<SIGNAL id=2>□ </SIGNAL>
<SIGNAL id=3>□□</SIGNAL>
<RLINK id=4 frame="VIEWER" signals="3"/>
<LINK id=5 source=1 target=3 linkType="EC"/>

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

18 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="US-NY" country="US"	GPE type="STATE-or-Province"
[Palm Beach] counties	PLACE type="CIVIL" state="US-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="CENTER"	Location
[Capitol Hill]	PLACE type="PPL" state="US-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>Moscow</PLACE> <PLACE type="COUNTRY" country="RU" id=2>Russia</PLACE> <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="TOP"	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" >town</PLACE> <SIGNAL id=2 type="DISTANCE">50 miles</SIGNAL> <SIGNAL id=3 type="DIRECTION">south</SIGNAL> <PLACE id=4 type="PPLA" country="AT" form="NAM">Salzburg</PLACE> <PLACE id =5 type="COUNTRY" country="AT" mod="CENTER">Austrian</PLACE> <PLACE id =6 type="MTS">Alps</PLACE> <RLINK id=7 distance=2 direction="S" source= 4 target=1 signals="2 3"/> <SIGNAL id=8>in</SIGNAL> <LINK id=9 source=1 target=6 signals="8" 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="BORDER">Thai</PLACE>	Relation: Part-Whole.GEO Location Arg1: [the Thai border] GPE Arg2: [Thai]
a military [base] [in] [Germany]	<PLACE type="FAC" id=1>base</PLACE> <PLACE type="COUNTRY" country="DE" id=2>Germany</PLACE> <SIGNAL id=3>in</SIGNAL> <LINK source=1 target=2 signals="3" linkType="IN"/>	Relation: Part-Whole.GEO FAC Arg1: [a military base in Germany] GPE Arg2: [Germany]
[St. Vartan's Cathedral], on [Second Avenue]	<PLACE type="FAC" id=1>St. Vartan's Cathedral</PLACE> <PLACE type="ROAD" id=2>Second Avenue</PLACE> <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>lobby</PLACE> <PLACE type="FAC" id=2>hotel</PLACE> <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>arena</PLACE> <PLACE type="FAC" id=2>Michigan State University</PLACE> <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

19 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.

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.

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

20 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).

21 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 type="DISTANCE">5 miles</SIGNAL>
<SIGNAL id=3 type="DIRECTION">east</SIGNAL>
<PLACE type="PPL" id=4 country="TW" form="NAM" latLong="2237N
12021E">Fengshan</PLACE>
```

<RLINK id=5 distance=2 direction=“E” source= 4 target=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.

22 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 type=“DISTANCE”>5 miles</SIGNAL>
<SIGNAL id=3 type=“DIRECTION”>east</SIGNAL>
<PLACE type=“PPL” id=4 country=“TW” form=“NAM”
latLong=”2262N 12034E”>Fengshan</PLACE>
<RLINK id=5 distance=2 direction=“E” source= 4 target=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.

23 Mapping to GUM

SpatialML has been mapped to the Generalized Upper Model (GUM) Ontology from the University of Bremen. The mapping of MODs is shown in Table 13.

MOD	Example	GUM Class
BOTTOM, TOP	the bottom of the [well]; the top of the [mountain]	VerticalProjection-Internal
CENTRAL	[central] Thailand; [central] Austrian Alps	CentralParthood; Distribution
E, N, ENE, ESE, NE, NNW, etc.	eastern [province], [North India], the [north] shore of Lake Lugano	CardinalDirectional-Internal
BORDER	[Burmese] border	Connection
NEAR	near [Harvard]	QualitativeDistanceProximal

Table 13: MOD Codes Mapped to GUM Classes

Note that the MODs can be separately mapped to binary relations by introducing an id for each MOD and relating it to the id for the tag being modified. For example:
the north shore of Lake Lugano[16]
<PLACE type="RGN" mod="N" id="1" form="NOM">shore</PLACE>
<PLACE type="WATER" id="2" form="NAM" latLong="45.967N 9.000E">Lake Lugano</PLACE>
<LINK source="1" target="2" linkType="EC"/>
In GUM, the corresponding relation could be represented as:
SpatialLocating (locatum "north shore",

16Example from Barker&Purves (2008). They treat "shore" as a geographic feature, whereas we treat it as a PLACE.

placement GL1 (hasSpatialModality NorthInternal, relatum "shore"))

Since such binary relations can be automatically derived, there is no need for reification of the modifier and separate LINKs between the modifier and the head. However, we will eventually need some indication of the extent of MODs.
Table 14 shows the mapping of Direction codes.

Direction	**Example**	**GUM Class**
BEHIND, FRONT	[behind] the house; [in front of] the theater	Horizontal Projection-External
ABOVE, BELOW	[above] the roof, over the clouds; [below] the tree-line, under the clouds	VerticalProjection-External
E, N, S, W, ESE, etc.	[E] of	CardinalDirectional-External

Table 14: Codes for Directions mapped to GUM Classes

Note that the Directions include external relations between entities, and some of the MODs are internal variants of those relations (e.g., *below the roof* versus *the bottom of the roof*).
In future, it may be suitable to add mappings for GUM classes such as LateralProjection-External, as in *to the left of the sofa*, *right of the church*, as well as for GeneralDirectional, such as *inside the house*, *outside the school*. These might correspond to LEFT (RIGHT) and INSIDE (OUTSIDE), respectively, in a future version of SpatialML.

24 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.

25 SpatialML Version

The "SpatialML" XML entity is a tag meant to contain the rest of the SpatialML content of a document, whether or not it contains the entirety of the document's content. It can act as the root tag for an XML document on its own, or it can be considered as just the root tag for the SpatialML subtree. The SpatialML tag has one attribute,"version", which is a string indicating the version of the XML standard being used.

26 SpatialML DTD (Version 3.0)

```
<!ELEMENT SpatialML ( #PCDATA | PLACE | RLINK | LINK | SIGNAL )* >
<!ATTLIST SpatialML xsi:noNamespaceSchemaLocation CDATA #IMPLIED >
<!ATTLIST SpatialML xmlns:xsi CDATA #IMPLIED >
<!ATTLIST SpatialML comment CDATA #IMPLIED >
<!ATTLIST SpatialML version 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 (
  BOTTOM | BORDER | CENTER | LEFT | NEAR | RIGHT | TOP |
  N | NNE | NE | ENE |
  E | ESE | SE | SSE |
  S | SSW | SW | WSW |
  W | WNW | NW | NNW ) #IMPLIED >
<!ATTLIST PLACE continent ( AF | AN | AI | AU | EU | GO | LA | NA | PA | SA )
#IMPLIED >
<!-- Country codes are ISO-3166 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 >
<!ATTLIST PLACE state NMTOKEN #IMPLIED >
<!ATTLIST PLACE latLong CDATA #IMPLIED >
<!ATTLIST PLACE nonLocUse ( true | false ) #IMPLIED >
<!ATTLIST PLACE predicative ( true | false ) #IMPLIED >
<!ATTLIST PLACE description CDATA #IMPLIED >

<!ELEMENT RLINK EMPTY >
<!ATTLIST RLINK id ID #REQUIRED >
<!ATTLIST RLINK comment CDATA #IMPLIED >
<!ATTLIST RLINK source IDREF #REQUIRED >
<!ATTLIST RLINK target IDREF #REQUIRED >
<!ATTLIST RLINK signals IDREFS #REQUIRED >
<!ATTLIST RLINK frame ( VIEWER | INTRINSIC | EXTRINSIC ) #IMPLIED >
<!ATTLIST RLINK direction ( BEHIND | ABOVE | BELOW | FRONT |
                      N | NNE | NE | ENE |
                      E | ESE | SE | SSE |
                      S | SSW | SW | WSW |
                      W | WNW | NW | NNW ) #IMPLIED >
<!ATTLIST RLINK 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 ) #IMPLIED >

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

27 Future Work

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

- Representing uncertainty.

References

Barker, E., and Purves, R. 2008. A Caption Annotation System for Georeferencing Images. Fifth Workshop on Geographic Information Retrieval (GIR'08). ACM 17th Conference on Information and Knowledge Management, Napa, CA, October 30, 2008.

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.