

Simple Named Entity Guidelines

Version 6.4 – Thai

March 28, 2006

Linguistic Data Consortium

Developed by Stephanie Strassel
(Based largely on the MUC-7 NE Guidelines)

Adapted to Thai by Kesarin Phanarangsarn, Sanipa Arnold, Mark Mandel
and Christopher Walker

1	Introduction	2
2	Entity Types	2
2.1	Person Names	3
2.2	Titles, roles and appositives.....	3
2.3	Organization Names	5
2.3.1	General ORGANIZATION-like non-entities.....	7
2.4	Location Names	7
2.4.1	Extent of Location Names.....	8
2.4.1.1	Compound expressions	8
2.4.1.2	Designators	9
2.4.1.3	Location modifiers and "semi-official" place names	9
2.5	Deciding among entity types.....	10
2.5.1	ORG referring to LOC, LOC referring to ORG	10
3	Difficult Cases	12
3.1	Expressions that refer to multiple entities	12
3.2	Nested Expressions	12
3.3	Entities as modifiers.....	13
3.4	Possessives (This rule does not apply to the Thai language.)	13
3.5	Other types of names.....	13
4	What NOT to tag	14
4.1	Events.....	14
4.2	Artifacts and products	14
4.3	Generics	14
5	Annotation Uncertainty.....	15

1 Introduction

An entity is some object in the world -- for instance, a place or a person. A named entity is a phrase that uniquely refers to that object by its proper name, acronym, nickname or abbreviation. Some examples of named entities follow:

Coca-Cola Bottling Co.	บริษัท ซี พี กรู๊ป จำกัด
Bob Austin	ธงไชย แมคอินไตย์
the Eiffel Tower	ตึกไอหยก 2
IBM	ปตท.
the Yankees	อิเหนา
Uganda	ญี่ปุ่น
Bowdon, Georgia	อ. สันกำแพง เชียงใหม่
Mt. Fuji	ดอยสุเทพ
the Kremlin	พระราชวังบางปะอิน
the Kennedys	ตระกูลเจียรนนท์

2 Entity Types

We will identify four types of named entities:

PERSON (PER): Person entities are limited to humans identified by name, nickname or alias.

TITLE/ROLE (TTL): Named personal titles or roles. These are restricted to titles that occur directly before or after the person name they describe.

ORGANIZATION (ORG): Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

LOCATION (LOC): Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

Other types of named entities like animals, inanimate objects and monetary units will not be annotated.

Within this document, named entities are indicated by [square brackets].

Note that annotations **may not** overlap or embed in the text. In other words, every annotation must end before another can begin.

2.1 Person Names

People may be specified by name, nickname or alias. Family names should also be tagged as PERSON. Names of deceased people, as well as fictional human characters appearing in movies, television, books and so on, should be tagged as PERSON entities. Religious deities should also be tagged as persons.

2.2 Titles, roles and appositives

Titles, roles and honorifics such as "Mr." and "President" are tagged as title entities and are separated from the individual's name. For instance, in the following sentences, there are two separate entities marked:

[Vice President] [Cheney] visited the site.

[นายกรัฐมนตรี] [ทักษิณ ชินวัตร] จะเปิดแถลงข่าวต่อสื่อมวลชน

For this task we define titles and roles as occurring either directly before or directly after a person name. Therefore, titles and roles are only tagged when they occur directly next to the person name they modify. In the following example, for instance, the phrases "Vice President," "นายกรัฐมนตรี" are not considered a title and are not tagged:

The strongest supporter was the Vice President.

นายกรัฐมนตรีผ่านร่างกฎหมายเกี่ยวกับการคุ้มครองสิ่งแวดล้อม

If a title contains within it a taggable entity, tag that entity separately¹. For instance:

[Microsoft] [Chairman] [Bill Gates] stated that...
ORG TTL PER

[ชาติรี โสภณพนิช] [ประธานกรรมการ] [ธนาคารกรุงเทพ จำกัด (มหาชน)]
ได้ปรับนโยบายบริหาร...
PER TTL ORG

(You may occasionally encounter an appositive like "Jr.", "Sr.", and "III". These are considered part of a person name and should be marked as part of the name. This rule does not apply in the Thai language.)

[Mr.] [Albert Franklin, Jr.] was part of the research team.
TTL PER

¹ Notice that the prohibition against overlapping or nested annotations (we may **never** insert one Named Entity *inside* of another one) forces us to turn the string "Microsoft Chairman Bill Gates" or "ชาติรี โสภณพนิชประธานกรรมการธนาคารกรุงเทพ จำกัด (มหาชน)" into a sequence of Named Entities. This is a special case for PERSON entities. In general, we may not break Named Entities into smaller pieces.

In Thai, the titles for royal family members are often used alone to refer to specific people. These titles should be marked as PERSON. However, if a title is followed by a specific name, mark the title as TITLE and mark the name as PERSON.

[พระบาทสมเด็จพระเจ้าอยู่หัว] เสด็จพระราชดำเนิน ไปทรงประกอบพิธีเปิดเขื่อน
PER

[พระบาทสมเด็จพระเจ้าอยู่หัว] (ภูมิพลอดุลยเดช)
ทรงเป็นพระมหากษัตริย์ที่ครองราชย์นานที่สุด
TTL PER

In Thai, the titles for monks are sometimes used alone without a specific name. These titles should thus be marked as PERSON. However, if a monk title is followed by a specific name –mostly in parentheses, the title will be marked as TITLE, and the name in parentheses will be marked as PERSON.

[สมเด็จพระพุฒาจารย์] เป็นปูชนียบุคคลที่พุทธศาสนิกชนให้ความเคารพอย่างสูง
PER

พระคาถาชินบัญชรโดย [สมเด็จพระพุฒาจารย์] ([โต พรหมรังสี])
TTL PER

In colloquial terms, Thai people often use titles, which indicate gender, seniority, etc. before their names. Some examples include คุณพ่อ, คุณครู, พี่, น้อง. These titles should be marked as TTL if followed by specific names. If they stand alone, don't tag them.

[พี่] [हनัน] โดนตัดสินพักงานทางการเมืองเป็นเวลา 5 ปี
TTL PER

[ก้านัน] [เป๊าะ] ถูกจับกุมในข้อหายักยอกทรัพย์สิน
TTL PER

พี่ยังคงเป็นที่เคารพของเหล่านักการเมืองรุ่นใหม่ no markup

ก้านันจะไปประชุมเรื่องการเลี้ยงปลา no markup

(Finally, sometimes the name of the person is split into two pieces by the title. In these cases, we will annotate the two pieces of the PERSON name as two separate PERSON entities. This rule does not apply in the Thai language.)

[Alfred] [Lord] [Tennyson]

PER TTL PER

Some more examples of names and titles:

[GlobalCorp] [Vice President] [John Smith]
ORG TTL PER

[Treasury] [Secretary] [Jackson]
ORG TTL PER

the [U.S.] [Vice President], [Dick Cheney]
 LOC TTL PER

[Justice] [Minister] [Giovanni Maria Flick]
ORG TTL PER

[British] [Rashtrodut] [Anwar Coudhury]
LOC TTL PER

[Mission Control] [Chief] [Vladimir Solovyov]
ORG TTL PER

In Thai, a person's name is usually located between their title and role. Since we are using TTL to tag both titles and roles, these terms will not be differentiated, but considered to be in the same category:

[นาย] [ทง พิทยะ] [รัฐมนตรีว่าการ] [กระทรวงการคลัง]
TTL (title) PER TTL (role) ORG

[ดร.] [คอนโดลิชซา ไรซ์] [รัฐมนตรี] [ต่างประเทศ] [สหรัฐ]
TTL (title) PER TTL (role) ORG LOC

[พล. ต. อ.] [เกา สารสิน] [รองประธานกรรมการ] [บมจ. ธนาคารกสิกรไทย]
TTL (title) PER TTL (role) ORG

2.3 Organization Names

Tag all proper name mentions of groups with a defined organizational structure. These include:

Businesses

[Bridgestone Sports Co.] profits

[บริษัท ยูนิลีเวอร์ ไทย เทรดิง จำกัด]

Stock exchanges

[NASDAQ] shares

[ตลาดหลักทรัพย์แห่งประเทศไทย]

Multinational organizations

[European Union] representatives

[อาเซียน]

Political parties

[GOP] hopeful

[พรรคไทยรักไทย]

Non-generic government entities

[the State Department]

[สำนักนายกรัฐมนตรี]

Sports teams

[the Phillies]

[ทีมการทำเรือ]

Military groups

[the Tamil Tigers]

[กองพันทหารม้ารักษาพระองค์ที่]

Many other kinds of entities refer to facilities or buildings that are primarily defined by their established organizational structure, and can do things like issue statements, make decisions, hire people, raise money and so on. A mention of such an entity should be tagged as an ORGANIZATION when it functions like an ORG in the document. These include things like:

Churches and other religious institutions

[Trinity Lutheran Church]

[วัดธรรมกาย]

Hospitals

[Finger Lakes Area Hospital Corp.]

[โรงพยาบาลพญาไท 1 จำกัด]

Hotels

[Four Seasons Hotel Group]

[โรงแรมเครือเซ็นทรัล]

Museums

[the Guggenheim Museum]

[พิพิธภัณฑ์สถานแห่งชาติ]

Universities

[the University of Chicago]

[จุฬาลงกรณ์มหาวิทยาลัย]

Government offices

[the White House]

[ทำเนียบรัฐบาล]

(Note that definite and indefinite determiners 'the' and 'a' are included in the annotation, except for cases when they quantify something other than the tagged entity, as in the following examples. This rule does not apply in the Thai language.)

A [Gulshan Hotel] spokesman

the [U.S.] Vice President

As in the above examples, this exception is particularly common when the tagged name is used in the pre-modifier (adjective) position.

In Thai, part of the organization name is sometimes followed by its abbreviation in parentheses, the whole name should not be broken up but marked together as ORG.

[คณะกรรมการธิการ (กมธ.) การต่างประเทศ]

ORG

[องค์การบริหารส่วนตำบล (อบต.) บ้านกลาง]

ORG

2.3.1 General ORGANIZATION-like non-entities

General entity mentions such as "the police", "the government", รัฐบาล, คณะรัฐมนตรี should not be tagged, since these are not unique proper name references to specific entities.

2.4 Location Names

Examples of place-related strings that are tagged as LOCATION include named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, highways, street names, factories, manufacturing plants, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, fictional or mythical locations, and monumental structures, such

as the Eiffel Tower, Washington Monument, อนุสาวรีย์ชัยสมรภูมิ, พระสมุทราชเจดีย์. For instance:

the collapse of the newly-constructed [Teton Dam]
LOC
the dispute over votes in [Dade County]
LOC
[The Walt Whitman Bridge] remained closed
LOC
repairs began on a 10-mile stretch of [the Alaskan Pipeline]
LOC
[The Garden State] is known for its tomatoes.
LOC

[โลก]
[เขื่อนแก่งกระจาน]
[เขตบางรัก]
[สะพานพระพุทธยอดฟ้า]

Note that in Thai journalism, country names are often referred to by nicknames and aliases. For example:

[ก๊ว]
[โสมขาว]

These nicknames and aliases should be marked as any other country name. They should be tagged according to their meaning and treated variably as ORG or as LOC.

2.4.1 Extent of Location Names

There are several issues surrounding the expression of location names and which parts of a string to tag.

2.4.1.1 Compound expressions

(Compound expressions in which place names are separated by a comma in English should be tagged as separate instances of LOCATION. This rule does not apply to the Thai language.)

[Kaohsiung], [Taiwan]
[Washington], [D.C.]

In Thai, compound expressions in which place names are separated by a space should be tagged as separate instances of LOCATION.

[อำเภอเมือง] [เชียงใหม่]
[ต. อินทรี] [อ. พรหมคีรี] [นครศรีธรรมราช]

2.4.1.2 Designators

When a "designator" is customarily used as a regular part of a place name, that word should also be included in the extent of the LOCATION entity. For example, include in the tagged string the word "River" in the name of a river, "Mountain" in the name of a mountain, "City" in the name of a city, etc., if such words are contained in the string.

```
[Mississippi River]
[the Himalayan Mountains]
```

```
[แม่น้ำเจ้าพระยา]
[ภูกระดึง]
```

2.4.1.3 Location modifiers and "semi-official" place names

Often times place names are modified by words like "Southern", "Lower", "West", "the former" and so on.

When these modifiers are part of a location's official name they should be tagged as part of the LOCATION name. For instance:

```
[Upper Volta]
[North Dakota]
```

```
[ภาคกลางตอนบน]
[ภาคกลางตอนล่าง]
[เยอรมนีตะวันออก]
```

Even if the place name does not have "official" status but has an agreed-upon definition and is in very frequent use, the string should be tagged as a LOCATION, as in:

```
[the Middle East]
[the West Bank]
[Eastern Europe]
```

```
[เอเชียตะวันออกเฉียงใต้]
[ตะวันออกกลาง]
[ยุโรปตะวันออก]
```

When these modifiers are not the official name of a place, or when the definition of the place might vary from person to person, do not tag the modifier as part of the LOCATION entity name.

```
[Mississippi River] west bank
former [Soviet Union]
[Gaul], in present-day [France]
lower [Manhattan]
Northern [California]
```

อดีต [สหภาพโซเวียต]
ปริมณฑล [กรุงเทพฯ]

These place names can sometimes be tricky. If you are not sure whether a modifier is part of an official name, you should include the modifier as part of the place name.

2.5 Deciding among entity types

There are some situations where deciding what entity type to assign can be somewhat tricky.

2.5.1 ORG referring to LOC, LOC referring to ORG

Many organizations have not only an organizational structure, but a physical location. For instance, museums are primarily organizations but are also housed in a specific building or facility. So while we normally tag museums as ORG entities, there are cases when a particular example might function more like a LOCATION. In cases like this, annotators should tag the named entity based on the way it functions in the sentence. For instance:

[The Guggenheim Museum] announced a new acquisition.
ORG

[The Guggenheim Museum] was designed by [Wright].
LOC PER

Thirty people were wounded in the bomb blast in front of the city's [Gulshan Hotel].
LOC

A [Gulshan Hotel] spokesman called the incident a tragedy.
ORG

[สวนสยาม] ทุ่มงบประมาณ 2 พันล้านปรับปรุงเครื่องเล่น
ORG

[สวนสยาม] เป็นสวนน้ำที่ใหญ่ที่สุดใน [เอเชียตะวันออกเฉียงใต้]
LOC LOC

ผู้ถือหุ้นของ [โรงพยาบาลพญาไท 2]
เข้าประชุมเพื่อรับฟังการแถลงผลการดำเนินงาน
ORG

มีการจัดกิจกรรมวันเด็กที่ [โรงพยาบาลพญาไท 2]
LOC

Similarly, city, country, and other place names are frequently used to refer to organizations located in those places rather than the geographical places themselves. For instance:

[Washington] announced a new tax policy today.
ORG

[โซล] เสนอตัวเป็นเจ้าภาพจัดการแข่งขันกีฬา
ORG

In this case, the name “Washington” and “โซล” is used to refer to the US Government, located in Washington and the South Korean Government, located in Seoul, respectively. Because “Washington” and “โซล” are referring to an organization entity in this example, it should be tagged as ORG.

Also, when the name of a unique structure or building (normally a location) is used to refer to the government or other organization housed in that facility, the name should be tagged as an ORG:

[The Pentagon] issued a statement about the incident.
ORG

[วังสราญรมย์] ออกประกาศแจ้งระเบียบการขอหนังสือเดินทางฉบับใหม่
ORG

The same logic applies to place names referring to sports teams. This rule does not apply to the Thai language.

[Boston] beat [New York] last night in extra innings.
ORG ORG

[จามจรี] เอาชนะ [แม็โดม] ในการแข่งขันฟุตบอลประเพณีประจำปี
ORG ORG

In summary, for any cases where a place name is used to refer to an organization, you should tag the name based on function in the sentence:

ORG: used when the example primarily refers to the organizational structure, and is acting like an agent (issuing a statement, making a decision, hiring people, raising money, etc.)

LOC: used when the example primarily refers to the physical structure, rather than the people/groups who run it.

3 Difficult Cases

3.1 Expressions that refer to multiple entities

When a phrase refers to multiple named entities, mark each entity separately.

For instance, this sentence contains two entities:

[China] and [South Korea] signed the agreement.

[จีน] และ [ไทย] ลงนามในสนธิสัญญาการค้าระหว่างประเทศ

Similarly,

[Jimmy] and [Rosalyn Carter]
[North] and [South America]

[วิลลี่] และ [แคทลียา แมคอินทอช]
[ยุโรปตะวันตก] และ [ตะวันออก]

But be careful not to split apart proper names that contain a conjunction.
For instance,

[the Fish and Wildlife Service]

[กระทรวงเกษตรและสหกรณ์]

is the name of one organization and should be tagged as a single named entity (it's not “the Fish Service” and “the Wildlife Service” and it's not “กระทรวงเกษตร” and “กระทรวงสหกรณ์” as separate names).

3.2 Nested Expressions

Recall that no nested expressions will be marked. When the name of one entity contains within it another entity name, do not pull out the name of the other entity and mark it separately². Only tag the larger entity. For instance

[Arthur Anderson Consulting] no markup for Arthur Anderson alone

[Boston Chicken Corp.] no markup for Boston alone

[กรุงเทพการบัญชี] no markup for กรุงเทพ alone

² PERSON entities that are adjacent to TTL and ORG pre-modifiers are a notable exception. Please see Section 2.2.

3.3 Entities as modifiers

If an entity name modifies another word (even if that word is not a taggable entity type), you should still tag the entity name.

[Bridgestone] profits
the [Clinton] government
[Treasury] bonds and securities
[U.S.] exporters
[Apple] computers
[Texas] intermediate crude oil
[China] film festival

รายการ [ทักษิณ] พบประชาชน
ชาว [อเมริกัน]
โครงการ [มียาซาว่า]
เครื่องเสียง [โซนี่]
อาหาร [จีน]
ผู้ส่งออกสินค้าการเกษตร [ไทย]

(Similarly, if the entity name occurs in the form of an adjective you should also tag it. This rule does not apply to the Thai language because Thai does not have an adjectival form.)

the [American] companies
[Cuban] citizens
[Chinese] food

3.4 Possessives (This rule does not apply to the Thai language.)

When you encounter a possessive construction, tag the two parts individually as two separate names. For instance:

[Temple University's] [Graduate School of Business]
[Canada's] [Parliament]

Keep in mind that annotation requires you to select whole words to tag as names, so you have to include the “s” even though it’s not part of the name.

3.5 Other types of names

Aliases, acronyms, nicknames and abbreviations for proper names should be tagged as a name:

IBM	[abbreviation for International Business Machines]
Big Blue	[alias for International Business Machines]
Big Board	[alias for New York Stock Exchange]
Mr. Fix-It	[nickname for candidate for head of the CIA]

the Big Apple	[nickname for New York City]
Red Sox	[alias for the Boston Red Sox]
Sears	[alias for Sears Roebuck and Co.]
มังกร	[nickname for the country 'China']
กีวี	[nickname for the country 'New Zealand']
กฟผ.	[abbreviation for
การไฟฟ้าฝ่ายผลิตแห่งประเทศไทย]	
โมเดิร์นไนน์ทีวี	[alias for
องค์การสื่อสารมวลชนแห่งประเทศไทย, อสมท.]	
บัวแก้ว, วังสราญรมย์	[nickname for กระทรวงการต่างประเทศ]
จามจุรี	[alias for จุฬาลงกรณ์มหาวิทยาลัย]
แม่โจม	[alias for มหาวิทยาลัยธรรมศาสตร์]

4 What NOT to tag

4.1 Events

Do not tag event names, even if they refer to events that occur on a regular basis and are associated with institutional structures. However, the institutional structures themselves — steering committees, etc. — should be tagged.

the Pan-American Games	[no markup]
vs.	
[the Olympic Committee]	[Organization]
โอลิมปิกเกมส์	[no markup]
[คณะกรรมการกีฬาโอลิมปิกเกมส์]	[organization]

4.2 Artifacts and products

Miscellaneous types of proper names that are not to be tagged as named entities include artifacts, other products, and plural names that do not identify a single, unique entity. For instance,

the Taurus is the latest car model	[no markup]
เปิดตัว [โซลูน่า] ในตลาดรถสำหรับขับในเมือง	[no markup]

4.3 Generics

Also, generic names that do not refer to a specific entity should not be tagged.

the Campbell Soups of the world	[no markup]
ก๋วยเตี๋ยวมาม่าในท้องตลาด	[no markup since 'มาม่า' here refers to all the brands of instant noodles]

5 Annotation Uncertainty

In some cases, you may encounter examples that you don't know how to handle. If so, you should proceed as follows:

- If it's an example not covered in the guidelines, note it in your copy of the guidelines and let your supervisor know about it.
- If it's an example where your language differs from the rules as written for English, note it in your copy of the guidelines and let your supervisor know about it.
- If it's a case where there's something wrong with the file you're working on, stop working on that file and let your supervisor know.

Your supervisor might not be available at the moment you have a question or issue, so it's important for you to write down the problem so it can be resolved later.

So that you can keep working even after you have a question about a particular example, we've created one more tag in the annotation tool: "No_Annotation". When you encounter a problem or have a question about a particular word or phrase and you can't get an immediate answer, label the item "No_Annotation". That will let us easily find it later when we try to resolve the problem.

You should also use the "No_Annotation" label in cases where there's some kind of problem with a single word or handful of words in the file – e.g., they're badly translated or the font isn't displaying properly.

If the whole file is problematic (i.e., poor translation, corrupted, font problems), stop working on it and let your supervisor know. If you are using AWS to receive file assignments, you may simply mark the file as "Broken" in the file-assignment interface. If you use this approach, please be sure to include a descriptive comment when AWS asks what is wrong with the file.