Annotation Guidelines for Individuality of Specific Entities

DEFT Entities, Relations, Events (ERE)

Version 2.6 – March 21, 2016

Linguistic Data Consortium
Created by: Ann Bies, bies@ldc.upenn.edu
With contributions from: Zhiyi Song, Joe Ellis, Neil Kuster, Justin Mott, Jeremy Getman, Xuansong Li,
{zhiyi,joellis,neilkus,jmott,jgetman,xuansong}@ldc.upenn.edu

Table of Contents

1 Introduction 3
2 Person Entities (PER) 4
3 Location Entities (LOC) 7
4 Facility Entities (FAC) 9
5 Geo-political Entities (GPE) 10
6 Organization Entities (ORG) 12
7 Appendix: Decision Tree Flowchart for Determining the Individuality of Specific Entities 13

1 Introduction

These guidelines are in addition to the existing ERE Entities annotation guidelines. Entities will be annotated as in the ERE Entities guidelines, with the addition of Individuality tagging as in these guidelines.

An addition to ERE annotation will be marking every specific (SPC) entity for whether the entity is

1. Individual (SPC_IND)
2. Group (SPC_GRP)
3. Unknown (SPC_UNK)

This distinction will be made at the Entity level, and will be inherited by all of the Entity's coreferenced entity mentions.

This distinction will be made only for specific (SPC) entities. It will be made for all Entity types: PER, LOC, FAC, ORG, GPE. Annotators will first decide whether an entity is specific or non-specific, exactly as is already the case in ERE annotation. The individuality distinction will be made only for entities that ERE annotates as specific (SPC).

This distinction must be made in the context of the document, as with all ERE annotation. For many instances, the context will help determine whether the entity is Individual or not.

The guiding principles are

1. For PER entities, is the reference to an individual person or not. This is really about whether the referent is one person or more than one person.
    a. One person → Individual
    b. More than one person → Group
    c. If you can't tell in context if it is one or more than one person → Unknown
    d. Other principles regarding names, etc. do not apply to this distinction for PER entities.
2. For LOC and FAC entities, there are two decisions, which should be made in this order:
    a. First, does the entity have a name (or is it likely or possible that it would have a name) in the world? This is really about whether the entity could have a proper name in the real world – it does not matter if that name actually appears in the document or not.
        i. Name → Individual
    b. Second, is the entity a unique, single location or facility in the world?
        i. Unique, single location or facility → Individual
        ii. Not a unique, single location or facility → Group
        iii. Can't tell in the context of the document if the entity is unique, single or not → Unknown
3. GPE and ORG follow the pattern of PER entities.

   a. One organization or geo-political entity → Individual
   b. More than one organization or geo-political entity → Group
   c. If you can't tell in context if it is one or more than one organization or geo-political entity → Unknown
   d. Other principles regarding names, etc. do not apply to this distinction for GPE or ORG entities.

4. If you cannot tell in context or from the guiding principles whether a specific entity is Individual or Group, mark it as Unknown.

More details and examples are in the sections below.

Note that the individuality distinction does not follow surface grammatical or morphological marking for singular or plural, and in many cases grammatically plural entities will be Individual while grammatically singular entities can be Group. The distinction is similar to the ACE PER subtypes of Individual, Group, Indeterminate. However, it is most closely related to the KBP and ED&L tasks, and is in support of ED&L. Entities marked SPC_IND in ERE will move into the annotation pipeline for the ED&L KB.

2 Person Entities (PER)

For PER entities, the only question is whether the entity refers to an individual person or not. This is really about whether the referent is one person or more than one person.

- If the entity refers to one person, mark it as Individual.
- If the entity refers to more than one person, mark it as Group.
- If you cannot tell in context whether the entity refers to only one person or to more than one person, mark it as Unknown.

The other principles in these guidelines (for non-PER entity types) regarding names, etc. do not apply to this distinction for PER entities.

Not Specific: it is either negated, irrealis or generic and therefore cannot be marked as a Specific Individual

```
either of the sisters
a typical voter
```

Individual:

```
the gunman
Barack Obama
Björk
Clinton
the speaker and audience members... [the entity with "speaker" as the head]
Hillary Clinton
Edmund Pope
The President of the U.S.
The police found [his] body
```

la esposa de Mandela
el portavoz del ministerio

[他]和[美国总统]的会谈颇有成效

Group:

the gunmen
The Obama family
the sisters
Billary [name for Bill and Hillary together]
the speaker and audience members... [the entity with “members” as the head]
Analysts
IBM’s lawyers
The family
The house painters
The linguists under the table
The Kennedys
The Arabs
The Christians
the Hmong individuals
the ethnic Albanians
multiple victims
the three perpetrators of a heist
“[Our project team for this assignment] is...” [PER Group because in context “team” does not have enough formal structure to be an ORG]
[The delegation] arrived yesterday [since “the delegation” is a PER entity consisting of more than one individual person (rather than an ORG), it is tagged as Group.]

Los Clintons [“the Clintons”]
La familia Mandela
mujeres americanas
la delagación china
un equipo médico (see note on “team” above)
los supervivientes del tifón Haiyan

[布什家族] [“Bush family”]
昨天[代表团]到达的时候 ... [when the delegation arrived yesterday...]
连日来，各地爆发大规模民众示威抗议，[示威者]与[警察]发生冲突，造成[数百人]受伤。

Unknown:

负责人在回答[记者]有关提问时作出上述表示的。
Al menos una persona fue herido

It is important to remember that the distinction between specific and non-specific must be made carefully first (see the Entities annotation guidelines for guidelines on specificity annotation), because the individuality distinction is only made for specific entities. The distinction of both SPC and NonSPC, as well as the distinction of Group, Individual or Unknown should be made within the context, as in these examples:

1. 有[人] 在爆炸中受伤。其中三人重伤
   `translation: Some people got injured in the bombing, and three are injured seriously.`

2. 有[人] 爆炸中受伤。该伤者送医后不治死亡。
   `translation: Someone got injured in the bombing. This injured person later died after being treated in the hospital.`

3. 有[人] 在爆炸中受伤。伤者被送医。
   `translation: Some people got injured in the bombing. The injured were transported to the hospital.`

人 (people) in all of the three examples above (or “some people” in English) is not generic, negated or irrealis in their contexts, and therefore it is not non-specific, but specific. Whether it is Group, Individual or Unknown, also depends completely in the context. In the first example, we know that there are more than three injured, so 人 is Group. In the second example, we know there is only one injured as it would be coreferenced with 该伤者, so it is Individual. In the third example, we don't know how many are injured and it can either be one or more than one, so 人 which would be coreferenced with 伤者, is Unknown.

Note that the category of PER Group includes family names and ethnic and religious groups that do not have a formal organization unifying them.

If from the context you cannot judge whether the Entity refers to an Individual entity or a Group of entities, tag it as **Unknown.**

It is also important to determine the Entity type of an entity in context first, since the Individual/Group distinction will be based in part on the Entity type. The same word could refer to different Entity types depending on the context, and as a result it would also have different individuality tags. For example, the word “team” could have the type PER or the type ORG, depending on its usage and referent in context. A PER “team” could refer to a pickup softball team one afternoon, for example – and the individuality for that entity would be Group PER. However, an ORG “team” could refer, for example, to a professional football team – and the individuality for that entity would be Individual ORG.

For the same reason, an example such as “Björk’s band” is not under consideration in this PER section – the band would probably not be in the PER category at all, because it should be marked as an ORG entity in most contexts (coreferent to “The Sugarcubes”, e.g.). If there is a context where such an example functions as a PER entity, it would be marked as Group, because it consists of more than one person. But as an ORG entity, it is an Individual entity, because it is one organization. See the ORG section for examples of this distinction for organizations.

3 Location Entities (LOC)

For LOC and FAC entities, there are two decisions, which should be made in this order:

- First, does the entity have a name (or is it likely or possible that it would have a name) in the world? This is really about whether the entity could have a proper name in the real world – it does not matter if that name actually appears in the document or not, and it does not matter if you can coreference a NAM mention with it in the document or not.
    - If the LOC or FAC entity has or could have a proper name in the real world, mark it as Individual.
    - If the entity would not have a name in the world, or if you can't tell whether it would or not, move on to the second decision point for LOC and FAC.

- Second, is the entity a unique, single location or facility in the world?
    - If the LOC or FAC entity is a unique, single location or facility, mark it as Individual.
    - If the LOC or FAC entity is not a unique, single location or facility, mark it as Group.
    - If you cannot tell in context whether the LOC or FAC entity is a unique, single location or facility or not, mark it as Unknown.

Individual:

the lake we went to
“When the [Hawaiian Islands], located far off the North American continent, were the site of...” [a reference such as this would likely be understood as a reference to the Hawaiian Island Chain as a location]
Rocky Mountains
Trinidad [the island of, rather than the country]
Lesser Antilles [a group of islands]
Mountains of East Kerry [name of a mountain chain or range]
Great Lakes

África
las Islas Senkaku
la área afectada

内赫布里底群岛 (Inner Hebrides, an archipelago/island chain)
五台山 (Five-Platform Mountain]
越南北部 (the northern part of Vietnam)
西方 (the West)

Group:

“visitors to the [islands of Hawaii] usually find that the ...” [only in a context where this is *not* referring to the Hawaiian Islands, which have a name]

“which is your favorite one of the Hawaiian Islands” [note that properly, “islands” should not be capitalized here, since this should not refer to the named entity]

America’s mountains

mountains of Bhutan

the rivers and streams in Pennsylvania

las zonas del conflicto de África

las islas del Mar Mediterráneo

中国有 [36 个西湖] Group，最著名的是杭州[西湖]Individual

Note that an example such as “Trinidad and Tobago“ would most likely be an Individual GPE entity, so it would most likely not be under consideration as a LOC entity. However, if there is a context where “Trinidad and Tobago“ refers to a location (the islands which have the same name) rather than the country, that each of it would be marked as a LOC individual.

It is important to remember that the individuality distinction has to be made in the context of the document. Let’s assume that the following examples (which include both grammatically singular entities and grammatically plural entities) are all SPC in context:

a large portion of northeastern Pennsylvania

most of the country

half the country

large portions of the country

la mayor parte de la provincia

la mitad del estado

algunas partes del país

江南 大部、华南 大部 有 中 到 大雨

translation: There will be moderate to heavy rain in most of Jiangnan and most of South China

Even so, we have to rely on the context to know if these SPC entities will be Individual or Group or Unknown:

- Individual (SPC_IND) if the context tells you they are contiguous places
- Group (SPC_GRP) if the context tells you they are separated, not contiguous places
- Unknown (SPC_UNK) if you can't tell from the context if they are together or not

In the example “There will be rain in most of the country (or large portions of the country)”, “most of the country” will be SPC_UNK because the context does not tell us whether the rain will be over a contiguous section of the country or over a number of separate places in the country. Similarly,

[江南 大部]Unknown、[华南 大部]Unknown 有 中 到 大雨

In the example “Temperatures in the south and north are in two extremes. Half the country is scorching and the other half is freezing”, “half the country” will be SPC_IND because it refers to a contiguous section of the country (either the South or the North).

4 Facility Entities (FAC)

For LOC and FAC entities, there are two decisions, which should be made in this order:

- First, does the entity have a name (or is it likely or possible that it would have a name) in the world? This is really about whether the entity could have a proper name in the real world – it does not matter if that name actually appears in the document or not, and it does not matter if you can coreference a NAM mention with it in the document or not.
    - If the LOC or FAC entity has or could have a proper name in the real world, mark it as Individual.
    - If the entity would not have a name in the world, or if you can’t tell whether it would or not, move on to the second decision point for LOC and FAC.

- Second, is the entity a unique, single location or facility in the world?
    - If the LOC or FAC entity is a unique, single location or facility, mark it as Individual.
    - If the LOC or FAC entity is not a unique, single location or facility, mark it as Group.
    - If you cannot tell in context whether the LOC or FAC entity is a unique, single location or facility or not, mark it as Unknown.

Individual:

Mall of America
Champs-Élysées
LAX
Six Flags Great Adventure & Wild Safari Kingdom
“I have to drive to [Philadelphia University] to pick up a friend”
“...the [university’s campus] includes...”
Houses of Parliament [proper name]
Rocky Mountain National Park
the highway
route 66
[the school outside Los Angeles]
University of Pennsylvania
the campus

el Centro Comercial Miraflores de la Ciudad de Guatemala
la Avenida Pensilvania

el Canal de Panamá
el nuevo megapuerto de Mariel

[这个十字路口]交通事故频发
[麒麟小区]的多栋高档独立别墅

Group:

malls in America
the boulevards of France
the buildings on campus
the houses on the street
the buildings at Penn

Centros comerciales guatemaltecos
Las avenidas de Pensilvania

[麒麟小区的多栋高档独立别墅]
中国政府去年建成[八条高速铁路]
[北京的所有机场]因大雾关闭

Even if the mention includes multiple buildings (like “the campus”), if the entity has a name (such as the “University of Pennsylvania”), we tag it as Individual.

Sometimes examples like this can go either way, and the distinction is based on whether the annotator interprets the entity in context as being an entity that is referring to a single place in the real world or not. For example,

[The massive complex] is made up of 10 buildings

If the massive complex that is referred to in the document is a complex that has a name in the real world, then this entity should be an Individual FAC. Or if the massive complex is referred to in context as a single place, then this entity should be an Individual FAC. Annotators should use their best judgment about the usage in context.

5 Geo-political Entities (GPE)

GPE entities follow the pattern of PER entities. That is, in most cases, it should be clear whether an entity refers to only one GPE or if it refers to more than one GPE.

- If the entity refers to one GPE, mark it as Individual.
- If the entity refers to more than one GPE, mark it as Group.
- If you cannot tell in context whether the entity refers to only one or to more than one GPE, mark it as Unknown.

Other principles regarding names, etc. do not apply to this distinction for GPE entities.

Individual:

Philadelphia
Bosnia and Herzegovina
Trinidad and Tobago
U.S.S.R.
United States
“when’re ya comin back to [the states] to visit?” [a mention of the U.S., and more usually capitalized]
American
Turks and Caicos
Los Angeles
[Miami] and [Tampa] [each entity in this coordination will be an Individual GPE]
[Tex]-[Mex] restaurant [two separate Individual GPE entities]
My country
Netherlands
European Union

la Unión Europea
la república ecuatoriana
los Estados Unidos

[波斯尼亚和黑塞哥维那] [Bosnia and Herzegovina]
“我决定离开 [这个让我迷失自己的城市]...”

Group:

“all of the **Philadelphias** in America”
“the Soviet Union countries”
“not all of **the states** are done voting”
The [Axis] powers in WWII

los países europeos
las nuevas repúblicas latinoamericanas
los estados norteamericanos
el bando beligerante que luchaba contra los Aliados

[欧盟创始成员国]
二战[轴心国]和 [同盟国]

If the GPE entity refers to a single geo-political entity, tag it as Individual. Even if the mention is grammatically plural (like “United States”), and even if it triggers plural verb agreement, if the entity is a single GPE, we tag it as Individual.

However, if the entity refers to a group of GPEs rather than a unique GPE (for example, if the entity itself does not have a government), we tag it as Group, even if it has a name or is grammatically singular (such as the WWII “Axis”). Note that GPE entities like the Unites States, U.S.S.R. or European Union are Individual GPEs because these entities themselves meet the GPE definition (they have a single government, etc.; see the ERE Entities annotation guidelines), even though they also contain sub-parts that are GPEs.

6 Organization Entities (ORG)

ORG entities follow the pattern of PER entities. That is, in most cases, it should be clear whether an entity refers to only one organization or if it refers to more than one organization.

- If the entity refers to one organization, mark it as Individual.
- If the entity refers to more than one organization, mark it as Group.
- If you cannot tell in context whether the entity refers to only one or to more than one organization, mark it as Unknown.

Other principles regarding names, etc. do not apply to this distinction for ORG entities.

Individual:

“when [that year’s team] won the Stanley Cup...”

Philadelphia Eagles

Pew Charitable Trusts

Philadelphia University

Björk’s band [when it is used as an ORG (not PER) in context]

“... the progressive wars can never be won. The [Democrats] will always find new hostile territory to invade...” [most likely being used as shorthand to refer to the Democratic Party as an Individual ORG rather than some subset of Democratic politicians or voters - see negative examples for Group PER below]

University of Pennsylvania

Red Cross

the soccer team

[Ford] and [Chrysler]

International Brotherhood of Teamsters

The [Democrats]

The [Republicans]

AARP

American Association of Retired Persons

[The winning team] accepted its trophy [Note that “team” is tagged as Individual because it is an individual organization, even though it consists of many people (it is not a PER entity).]

la Unión de Mujeres Americanas

Ministerio de Seguridad de la Nación

el gobierno brazilero

[阿里巴巴集团] [Alibaba Group]

Group:

Philadelphia’s sports teams

[Pew’s charities]

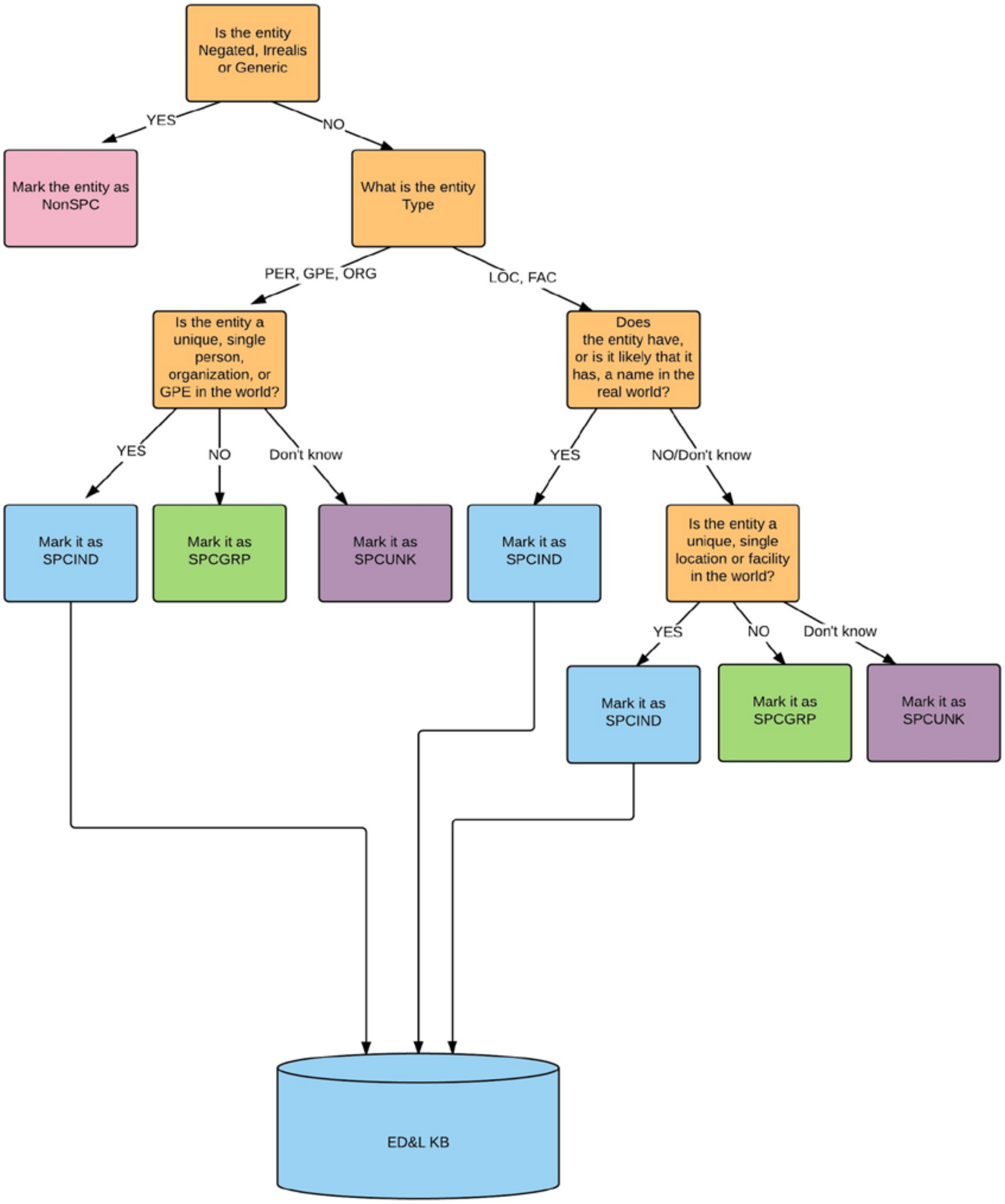
the two houses of Congress

“if you believe what Democrats say about...” [vague reference to people of Democratic leaning]

“House Democrats today announced that...” [multiple people , not an organization]

[The NGOs] in the area

7 Appendix: Decision Tree Flowchart for Determining the Individuality of Specific Entities


Is the entity Negated, Irrealis or Generic
YES
NO
Mark the entity as NonSPC
What is the entity Type
PER, GPE, ORG
LOC, FAC
Is the entity a unique, single person, organization, or GPE in the world?
Does the entity have, or is it likely that it has, a name in the real world?
YES
NO
Don't know
YES
NO/Don't know
Mark it as SPCIND
Mark it as SPCGRP
Mark it as SPCUNK
Mark it as SPCIND
Is the entity a unique, single location or facility in the world?
YES
NO
Don't know
Mark it as SPCIND
Mark it as SPCGRP
Mark it as SPCUNK
ED&L KB