RUFES 2022 Annotation Guidelines (V1.1)

Change logs:

- 2021: major changes from 2020
  - Entity types expanded
  - Expand generic reference (for example, tag predicate NP as generic)
  - Treat hyphenated compounds the same way as closed compounds
  - Negative mentions are potentially taggable
- 2022: changes from 2021
  - Remvoved PER.CivilServant.Spokesperson but added PER.Professional.Spokesperson. So for a government spokesperson, they must be labeled as PER.CivilServant as well
  - Changed ORG.CommercialOrganization.Carrier to ORG.CommercialOrganization. TransportCompany
  - Changed VotingFacility with PollingPlace and augmented it to level 2 directly under FAC
  - Added 45 new entity types (see the new ontology spreadsheet for details)
  - Added Chinese Supplement

1. Introduction

RUFES (**R**ecognizing **U**ltra **F**ine-grained **E**ntitie**s**) is a continuation of Entity Discovery and Linking (EDL) tasks, with focus on fine-grained (broader and deeper) entity types. While these annotation guidelines are developed primarily for NIST annotators, participants are encouraged to study them in conjunction with task specifications to ensure new features and requirements are implemented.

A sample Washington Post article is included in this document and references will be made to it for illustration. A table showing annotation output for this sample document can be found in Appendix II.

1.1 Annotation Objective

An EDL system under RUFES automatically detects entity mentions in a news article that refer to entities as defined in the ontology, assigns one or more entity types that can be determined from the context, and links mentions that refer to the same entity within the document. Annotators will perform the same task to create a gold reference on which system output can be evaluated for performance. Annotators must strive to achieve the highest possible accuracy. For quality assurance, a second pass will be performed on each document by a different annotator.

1.2 Basic Concepts

- ***Entity*** – a physical object or a conceptual reality that exists in some universe (typically the real world in which we live).
- ***Word*** (for written English) – a meaningful sequence of letters between two spaces or between a space and a punctuation.
- ***Phrase*** – one or more words that function as a grammatical unit in a sentence (such as the subject, object, etc.).
- ***Head*** – the word in a phrase that determines the syntactic category of the phase.
- ***Mention*** – a phrase that refers to an entity.

- ***Mention extent*** – the taggable part of a mention.
- ***Reference*** – a relationship between a mention and an entity, such that the mention designates the entity.
- ***Coreference/Linking*** – a relationship between two or more mentions such that they refer to the same entity.
- ***Ontology*** – a model that organizes structured and unstructured information through entities, their properties, and the way they relate to one another.

What mentions refer to what entities in the first two paragraphs of the sample news article? Is there coreference?

> As James Comey takes over as the new FBI director, the American Civil Liberties Union is calling on the Obama administration and Congress to rein in the increasing power of the agency.
>
> In a critical 63-page report that will be issued Tuesday, the ACLU says the powers of the FBI have expanded too dramatically over the past 12 years, transforming the Bureau into a "secret domestic intelligence agency."

1.3 Taggability

A mention is taggable if and only if it refers to an entity that is an instance of one of the predefined entity types. Consider the third paragraph from our sample article.

> "The excessive secrecy with which it cloaks these domestic intelligence gathering operations has crippled constitutional oversight mechanisms," the report says. "Courts have been reticent to challenge government secrecy demands and, despite years of debate in Congress . . . it took unauthorized leaks by a whistleblower to finally reveal the government's secret interpretation of these laws and the Orwellian scope of its domestic surveillance programs."

There are several examples where the NP refers to something that's not covered by the RUFES ontology, for example, "the excessive secrecy" (referring to a condition), "mechanism" (referring to some systems), etc. Taggable mentions in this paragraph include "report", "courts", "Congress", etc.

1.4 Reference

Reference is a relationship between a mention and an entity. Philosophers and linguists recognize different kinds of reference. Reference can be definite/indefinite, new/given, and singular/plural.

- John is driving my car.
- John is driving a car.
- Lindsey Graham stormed out of the room.
- The lawmaker stormed out of the room.
- A lawmaker stormed out of the room.
- Two lawmakers are opposed to the bill.
- Some lawmakers are opposed to the bill.
- The two lawmakers are opposed to the bill.

Reference can also be specific or generic. A mention makes specific reference if it refers to one or more specific entities. Mentions above are all examples of specific reference.

If the reference is about a class of entities, it's known as generic reference.

- `Politicians only care about their constituents.`
- `A politician only cares about their constituents.`
- `The politician only cares about their constituents.`

Do not confuse specificness with definiteness. The last example is an NP with a definite article but makes generic reference to politicians.

Change from 2020!

For the 2020 annotation, a distinction was made between specificational and predicational NPs as shown in the following pair of sentences:

- `Barack Obama was the 44th US president.`
- `Barack Obama is a former US president.`

The phase "the 44th US president" is known as specificational because it refers to a specific president in the US history whereas the phrase "a former US president" ascribes some property to the individual that the subject NP "Barack Obama" refers to, namely, he was a US president at some point in the past. Predicational NPs are considered non-referential in some linguistics theories and RUFES 2020 adopted that approach.

Some recent entity tasks (e.g. AIDA) are interpreting predicational NPs as generic reference. RUFES 2021 will follow suit. Note though, "the 44th US president" and "Barack Obama" in the first sentence refer to the same entity. But "a former US president" in the second sentence does not refer to the entity named "Barack Obama". "Being a former US president" means there is a set of people who have served as US presidents and "Barack Obama" is a member of that set. In other words, "Barack Obama" is an instance of the class "former US presidents".

When an unspecific mention is within the scope of a negation, we will also treat the reference as generic.

- `The US and its coalition did not find [weapons] of mass destruction in Iraq.`

There's a generic reference of a predicational NP in the second paragraph of the sample article. Can you identify it?

While we do not make such distinctions in the annotation output, knowing them will be helpful in understanding and performing the task, especially when annotating coreference.

2. Mention Types and Mention Extent Selection

We distinguish between three types of mentions – named mention, nominal mention, and pronominal mention. As we define and illustrate each mention type below, we will also show how the mention

extent (the beginning and end boundaries) should be annotated for each mention type. It is very crucial to select the correct boundaries.

2.1 Named Mention (NAM)

A named mention (NAM) is a noun phrase whose head is a proper name, or any of its variants such as acronym, nickname, abbreviation, alias, etc. that refers to an entity. In the above example paragraphs, *James Comey*, *the FBI*, *the American Civil Liberties Union*, *the ACLU*, *Obama*, and *Congress* are all NAMs. Note proper adjectives, namely adjectives derived from proper nouns, are also taggable (details below).

2.1.1 NAM Extent

The taggable extent of a NAM is the entire proper name excluding any determiner (except where the determiner is always part of the name), or any modifier that appears before or after the proper name (premodifier or postmodifier) that is not an integral part of the name.

- `the FBI -> the [FBI], ~~[the FBI]~~`
- `James Comey -> [James Comey], ~~[James] [Comey]~~`
- `the ACLU 's Washington Legislative Office -> the [ACLU] 's [Washington Legislative Office]`

NAMs are atomic and no additional mention should be tagged within the extent even if some words may seem to be mentions of other entities.

- `the University of Maryland -> the [University of Maryland], ~~the [University of [Maryland]]~~, ~~the [University] of [Maryland]~~`
- `the United States Department of State -> the [United States Department of State], ~~the [[United States] Department of State]~~, ~~the [United States] [Department of State]~~`

But

- `Georgetown University in D.C. -> [Georgetown University] in [D.C.]`
- `University of Oregon in Eugene, Oregon -> [University of Oregon] in [Eugene] , [Oregon]`

Compare also:

- `the Chinese Foreign Ministry -> the [Chinese Foreign Ministry]`
- `China's Foreign Ministry -> [China]'s [Foreign Ministry]`

Note, mention nesting - a mention extent is embedded within another mention extent - is not supported under RUFES. Annotators need not worry as the annotation tool does not permit mention nesting. This does not apply to nominal mentions with taggable modifiers as the taggable extent of a nominal mention is the head word only.

However, the non-compositional aspect of NAMs is only applicable to taggable mentions, meaning the entity type is in the list of entity types. If the entity type is not covered by the current ontology, then a proper noun appearing in the proper name may be taggable if the entity it refers to is an instance of a supported entity type.

- `Tamerlan Tsarnaev, one of the [Boston] Marathon bombing suspects`

2.1.2 Proper Adjectives

In English, certain proper nouns such as those referring to geological entities such as countries, cities, regions, mountains, etc. have adjectival forms known as proper adjectives, for example, *America* -> *American*, *China* -> *Chinese*, *Himalayas* -> *Himalayan*, and so on. These adjectives are taggable and will be treated the same as NAM.

- `the [American] [Ambassador] to [Japan]`
- `the [Chinese] [Ambassador] to [Japan]`
- `[Russian] – made`
- `[Asian] American`

2.1.3 Special considerations

Personal names may have titles, professional credentials, etc. They are not integral parts of personal names and thus must be excluded from the mention extent:

- `President [Joe Biden]`
- `Dr. [Anthony Fauci]`
- `[Jane Doe], M.D.`

Not, though, when a tile is used directly to address a person, it is taggable (but as a nominal mention). For example,

- `[Doctor], I have a question.`
- `[Colonel], permission to speak.`
- `Mr. [President], are you going to run in the election?`

Generational suffixes such as Jr., Sr, II, etc. suffixes are an integral part of personal names and must be included in the mention extent.

- `Donald Trump Jr. -> [Donald Trump Jr.],` ~~`[Donald Trump] Jr.`~~

Note to annotators: the text displayed by the annotation tool does not look the same as you would normally read an article online or on paper. The text has been tokenized such that (among other things) punctuations are usually separated with a space from the neighboring words. For example, a name such as *O'Brien* would be rendered as `O ' Brien`. Your selection must include all the tokens in its extent, including punctuations that are part of a name: `[O ' Brien], the [U . S .]`. However tokenization keeps a sequence of punctuations as a single string, for example `"U . S .- led war"`. In such cases, do not include the punctuation cluster at the end of the mention.

- `[U . S] .- led war`

Sometimes a determiner is an integral part of a name, for example, *The* in *The Limited* (a clothing company that is no longer in business) or in *The Hague* (a European city). In such special cases, the determiner must be included in the mention selection.

Note a passenger airplane may also be referenced by a flight number even though different airplanes may be flown on different days for the same flight number (or the same airplane may assume different flight numbers). We currently do not have flights in the ontology. But if an airplane is referenced by the

flight number, the flight number (for example, the ill-fated MH370) will be tread as a NAM for the airplane.

2.2 Nominal Mention (NOM)

A nominal mention is a noun phrase headed by a common noun. In the first paragraph of our sample article, there are two nominal mentions: *the new FBI director* and *the Obama administration*.

2.2.1 NOM Extent

For nominal mentions, only the head of the noun phrase is selected to represent the mention.

- `In its [report], the ACLU asks Congress, the [president] and the [attorney] general to conduct a comprehensive evaluation of the FBI's policies and programs, and makes 15 recommendations for reform of the [agency].`

We previously mentioned that a tagged mention cannot be imbedded within another tagged mention. However, because we only select the head of a NOM, taggable mentions within the full boundaries of a NOM can and must be annotated.

- `the new FBI director -> the new [FBI]`$_{NAM}$ `[director]`$_{NOM}$
- `the Obama administration -> the [Obama]`$_{NAM}$ `[administration]`$_{NOM}$

2.2.2 Compound Noun as NOM Head

English distinguishes between three written styles of compound nouns:

- Closed (solid) compounds: *marketplace, supermarket, cowboy, firefighter*
- Hyphenated compounds: *editor-in-chief, court-martial, hunter-gatherer, bar-restaurant*
- Open compounds: *vice president, assistant professor, attending nurse, chief justice, attorney general, editor in chief*

A closed compound noun is always treated as one word (the annotation tool does not allow partial word selection). For open compounds, only the head is selected as the mention extent. This primarily because it is not always easy to distinguish between an open compound and a noun with a modifier.

- `vice [president]`
- `assistant [professor]`
- `attending [nurse]`
- `chief [justice]`
- `[attorney] general`
- `[editor] in chief`

As shown in the last two examples, the head of a compound in English is not always the final word.

In RUFES 2020, hyphenated compounds were treated the same way as open compounds. For 2021, we are treating them as "atomic" words instead.

- `[editor - in - chief]`
- `[father - in - law]`

This reversal is made in part because "headless" compounds tend to be written with hyphens.

- [hunter – gatherer]
- [bar – restaurant]
- [tractor – trailer]
- [writer – editor]

2.2.3 Nominalized Adjective

Some adjectives can be used directly as the head of a noun phrase, for example, *the rich*, *the poor*, etc. In such cases, we select the adjective as the mention head.

2.2.4 Number as Mention Head

Numbers can function as the head of an NP when the "missing" noun appears in a previous clause or sentence:

- There were five [survivors]; [two] were children.

They can also be used as the head in a partial-whole construction:

- [One] of the five [survivors] was a child.

Note the above example has two mentions, one with a number as the head and one with a common noun as the head.

If the number consists of more than one word, the entire sequence of words must be tagged as the mention head. This also applies to percentage or fraction.

- Of the 7,160 patients whose chronic illness status was known through health records, [184] died, and [173] of them had an underlying condition, the CDC said.
- Among all the cases analyzed, [10.9 percent] of patients had diabetes mellitus, [9.2 percent] had chronic lung disease, and [9 percent] had cardiovascular disease.

Note "10.9 percent of patients" has two mentions, [10.9 percent], [patients], where [patients] refer to all the patients under the study and [10.9 perfect] refers to a portion of them.

2.2.5 Possessive NP without a Head Noun

When a head noun follows a possessive noun phase, for example, "car" as in "John's car", it may be omitted when the noun appears in a previous phrase/clause/sentence. In this case, we select "s" as the mention head:

- John ' s [car] is black; his wife ' [s] is red.

If the possessive NP is a plural that ends in "s", the apostrophe is selected as the mention head.

- John ' s house is next to his parents ['].

2.2.6 Pro-noun (Not Pronoun!)

In English, given the right context, the word *one* and its plural form *ones* can be used in place of a common noun, for example, *the green one*, *the poor ones*, *the one you picked out*, etc. In such cases, we select the pro-noun as the mention head. Note the difference between prop-words/pro-nouns and pronouns. The word *one* can be used as a pronoun as well, as in "One should not steal". But pro-nouns

can have premodifiers and determines just like common nouns. Pronouns on the other hand do not usually take premodifiers. We must be careful to determine the correct mention type.

2.2.7 Proper Noun as NOM Head

A noun phrase headed by proper noun isn't always a NAM mention.

- John talked to a [Guatemalan]$_{NOM}$ attempting to immigrate to the U.S.

But,

- John met a [Guatemalan]NAM [immigrant]NOM yesterday.

Another common usage of proper nouns as NOMs is a manufacture's name used to refer to its product(s) or a product model name used to refer to an instance of that model.

- John drives a Toyota.
- My Honda broke down yesterday.
- John drives a Corolla.

In such as case, the proper noun is labeled as NOM as well because the manufacturer name or the model name is not the name for the vehicle.

2.2.8 Not taggable: Singular Bare Noun as Modifier or in Idiomatic Expressions

A singular bare noun that modifies another noun or appears in some idiomatic expressions is not a noun phrase and should NOT be tagged:

- ~~car~~ manufacturers, ~~furniture~~ maker, …
- go to ~~school~~, in ~~hospital~~, …

Some idioms can no longer be interpreted literally.

- kick the bucket, …

2.3 Pronominal Mention (PRO)

A pronominal mention is a noun phase consisting of a pronoun. In terms of how pronouns make reference to entities, they may be classified as deictic (type of exophoric referring to something in the context of the speaker), or endophoric (referring to something mentioned in the surrounding text), which can be either anaphoric (backward reference) or cataphoric (forward reference).

- (*Speaker pointing to John*) He is Mary's brother.
- John walked inside. He is Mary's brother.
- When he walked inside, John ran into his sister.

The interpretation of a pronoun is context dependent. So, whether a pronoun is taggable depends on the entity it refers. Most pronouns are potentially taggable except for interrogative pronouns.

2.3.1 Taggable and Non-Taggable PRO

Here is a list of potentially taggable pronouns:

- Personal pronouns
  - Subjective: *I, we, you, he, she, they, it*

    - Objective: *me, us, you, him, her, them, it*
    - Possessive
        - Possessive determiners: *my, our, your, his, her, their, its*
        - Possessive pronouns: *mine, ours, yours, his, hers, theirs, its* (but be careful about the entity they refer to!)
- Demonstrative pronouns: *this, that, these, those*
- Reflexive pronouns: *myself, ourselves, yourself, yourselves, himself, herself, themselves, itself, oneself*
- Reciprocal pronouns: *each other*, *one another* (both words are selected as the head)
- Relative pronouns: *who*, *whom*, *whose*, *which*, *that, whoever, whomever*
- Dual pronouns: *both, either, neither*
- Positive indefinite pronouns: *all, some, one, many, everything, everybody*
- Negative indefinite pronouns: *no one, nobody, nothing, nowhere* (context-dependent)

We also extend pronominals to proadverbs such as (e.g. *there*) and relative adverbs (e.g. *where*).

- `John went to the library. He did not borrow any books [there].`
- `John went to the library yesterday, [where] he met an old friend from college.`

A departure from RUFES 2020 is that negative pronouns (e.g. *no one, neither, etc.*) may be taggable if the context allows it. Compare the following two examples.

- `Although Jenkins and Rankovicz did well in the primaries, neither won a majority of the vote.`
- `Although Jenkins and Rankovicz did well in the primaries, both failed to win a majority of the vote.`

The two sentences have pretty much the same meaning. We will tag "neither" in the first example the same way as we tag "both" in the second example.

Interrogative pronouns (used in questions) such as *who*, *whom*, *whose*, *what*, *which* are not taggable.

Note that a relative pronoun is not taggable if the antecedent is nonreferential:

- `I have a new doctor, [who] is very patient.`
- `I bought a book last week, [which] I haven't read.`
- `My sister's doctor is very patient, ~~[which]~~ my doctor isn't.`

2.3.2 PRO Extent

The extent to be selected for a taggable pronominal mention is simply the pronoun itself. Pronouns usually don't take premodifiers, though some (usually for generic reference) may have a relative clause as a postmodifier, which will not be included in the extent selection.

- `[He] [who] can no longer pause to wonder and stand rapt in awe is as good as dead.`
- `[Those] [who] do not complain are never pitied.`

2.4 Conjoined Mentions

When two or more mentions are conjoined, they must be tagged separately:

- `Boeing and Lockheed Martin -> [Boeing] and [Lockheed Martin]`

Mention selection must be continuous. We cannot skip a word and select two discontinuous words to form a single mention.

- `Barack and Michelle Obama -> [Barack] and [Michelle Obama]`

In this case, we will only select *Barack* as a mention referring to the former president.

Although conjoined mentions function as a single grammatical unit such as the subject or object of a sentence, we do not create another mention for the entire conjunction.

- `Boeing and Lockheed Martin -> ~~[[Boeing] and [Lockheed Martin]]~~`
- `Barack and Michelle Obama -> ~~[[Barack] and [Michelle Obama]]~~`

2.5 Apposition

An apposition is a construction where one (or more) noun phrase follows another noun phrase to provide more information about the first noun phase:

- `My sister, Alice Smith, likes jellybeans.`
- `Alice Smith, my sister,  likes jellybeans.`
- `My sister, a doctor, likes jellybeans.`

Both mentions in an apposition may be taggable. Challenge: what does *a doctor* refer to?

2.6 "False" Titles

There is this notion of "false titles" that is widely used in (American) journalist writings. They look like titles (which are not taggable as mentioned earlier) but they are not capitalized, and their meaning is more like a preposed appositive:

- `cellist Yo-Yo Ma`
- `famed New Left philosopher Herbert Marcuse`
- `convicted bomber Timothy McVeigh`

We will treat them the same way as we do with appositions. But because they tend to omit determiners, their reference can be harder to determine. We'll discuss this in section 4.

3. Assigning Entity Type

Our mind perceives the world structurally, and so part of the goal of an ontology is to categorize entities in a hierarchical structure. For example, there are cars (big or small), vans, SUV's, trucks, etc., all of which are propelled by a motor and do not run rails. So, we group them as motor vehicles. Then there are all kinds of trains that run on rails. So, we say they are railed vehicles. We also have bicycles, tricycles, scooters, and they are powered by our own muscle. We call them human powered vehicles. All motor vehicles, trains and human powered vehicles run on wheels. So, together we call them wheeled vehicles. Wheeled vehicles, along with boats, ships, etc. that run in water (which we call watercrafts), and airplanes and helicopters that fly in the air (which we call aircrafts), are used to transport people or cargo. We simply call them vehicles.

That is how we categorize entities – based on their commonalities and by levels. Previous entity detection tasks mainly focused on top level entity types. Those "coarse" entity types have become less and less sufficient for entity detection and other entity related technologies. So, entity detection is moving on to more and more refined entity types.

The RUFES entity set organizes entity types into three levels. In our vehicle example, vehicle is a level 1 entity type, wheeled vehicle is a level 2 entity type, and railed vehicle is at level 3.

RUFES 2021 has a total of 309 (from 266 in RUFES 2020) entity types: 15 at the very top (introducing MED for medication), 94 at the second, fine-grained level (an increase by 16 from RUFES 2020), and 199 at the third, ultra -fine-grained level (an increase by 27 from RUFES 2020). For a complete list of entity types under RUFES 2021, refer to Appendix III or the accompanying ontology spreadsheet.

Below are entities from the first paragraph of the sample document.

```
As [James Comey] takes over as the new [FBI] [director], the [American
Civil Liberties Union] is calling on the [Obama] [administration] and
[Congress] to rein in the increasing power of the [agency].
```

- `[James Comey], the new FBI [director]: PER.CivilServant`
- `[FBI], the [agency]: ORG.Government.Agency`
- `The [American Civil Liberties Union]: ORG.AdvocacyGroup`
- `the Obama [administration]: ORG.Government.Administration`
- `[Obama]: PER.Politician.HeadOfGovernment`
- `[Congress]: ORG.Government.Legislature`

There are two closely related restrictions on when we decide what entity type label(s) to assign for an entity mention: context and usage - we only tag a mention if the entity it refers to falls into one of the entity types in the ontology as can be determined by where it occurs and how it is being used.

3.1 Scope of Context

Imagine we are annotating George H.W. Bush's obituary. He served as President, Vice President, ambassador to US, top diplomat to China, congressman, WWII veteran, and so on and so forth in his lifetime according to the obituary. Does it mean we should tag every mention of him in the article with all these entity types? The answer is no. Instead, for a given mention, we only label it with entity types as can be determined from the paragraph in which it occurs. We do not "inherit" entity types that are assigned to mentions of the entity that appear in previous paragraphs, nor do we read beyond the current paragraph to find entity types applicable to later mentions that refer to the same entity.

Just as a news reporter expects the reader to have some level of world knowledge (for example, an American reader should know what "FBI" refers to), annotation also assumes some level of world knowledge on the part of the annotator (most of us indeed know that the FBI is a US government agency). However, that does not mean we should label a mention with all entity types we know. For example, if you happen to have read James Comey's bio and know he was a lawyer in his early carrier and was also a business executive at some point, you should not assign PER.Professional.Lawyer or PER.Executive.BusinessExecutive to any mention of James Comey in the sample document because neither entity type is explicitly mentioned or implied in the article.

Consider the opening paragraph from a news report about Amy Coney Barrett's Supreme Court Justice nomination hearings below:

- Senate Republicans moved swiftly Thursday toward confirming Supreme Court nominee [Amy Coney Barrett] before the Nov. 3 elections, as the hearings for President Trump's third pick to the court concluded with her emerging largely unscathed.

What entity type should we assign to the mention "Amy Coney Barrett"? If we followed the news, we knew she had been a circuit judge since 2017. So we might be tempted to tag it as PER.CivilServant.Judge. But the paragraph itself does not mention this information. If we are familiar with qualifications of a Supreme Court Justice nominee, being a current judge or even a lawyer is not one of them. Therefore, we should only label it as PER in this paragraph (if we had something like PER.PoliticalNominee as an entity type, that would be a more appropriate label).

If we keep on reading, in a later paragraph, we see a mention of her as "Judge Barrett". So in this paragraph, PER.CivilServant.Judge should be assigned.

- "Judge [Barrett] as Justice [Barrett] may well cast the deciding vote to overturn the Affordable Care Act, with potentially disastrous consequences for a majority of Americans," Sen. Christopher A. Coons (D-Del.) said Thursday.

In general, we should avoid using information external to the news article to determine entity types except when you encounter an unfamiliar entity that you may need confirmation, for example, an unfamiliar location or landscape in a foreign country that you are not sure if it is a GPE or a LOC.

3.2 Tag for Usage/Meaning

While context defines the scope where we look for evidence to determine the entity type of a mention, the actual entity that the mention is intended to refer to is the label we should use. This has been know as the "tag for usage" or "tag for meaning" rule. In most cases, the entity a mention refers to is the one we tend to associate with in our mind. But that is not always the case.

For example, a very common figure of speech in journalist writings is metonymy, where another name closely associated with the entity is used instead of its official name, for example, the *Pentagon* referring to the US military, the *White House* referring to the US government executive branch, etc. In such cases, the entity type we assign to the mention is determined by the intended entity being referred to. For example, in "The White House just announced ….", we should label the *White House* as ORG.Government.Administration. On the other hand, in a sentence such as "The White House is currently not open to visitors", we should assign FAC.Building.GovernmentBuilding to the *White House*.

Other common examples of metonymy include the capital city's name referring to the country's or state's government, a city name referring to its sports team(s), etc.

Facilities are operated by organizations and often the same name may be used for both kinds of entities. For example, *McDonald's* may refer to the corporation (as in "McDonald's supports raising longtime $7.25 floor"), which is ORG.CommercialOrganization, or one of its restaurants around the corner (as in "Jonny bought a kid's meal at McDonald's"), which is FAC.Building.StoreShop (we currently do not have restaurant as an entity type). Other examples are hospitals, educational institutes, etc.

Sometimes, a generic reference to people is intended to refer to their organization or their country, for example, the *Democrats* referring to the *Democratic Party* (as in "The Democrats won the House"), or the *Americans* referring to the *United States* (as in "The Americans declared war on the Japanese on December 8, 1941"), or *(government) officials* referring to the government, etc.

How should we tag "Republicans" and the "they" in the second sentence?

- `If Republican Dan Sullivan holds his lead in the Alaska Senate race and Cassidy wins, Republicans will have gained nine Senate seats. They needed six to win the majority.`

3.3 Levels of Entity Types

In the RUFES ontology hierarchy, every top-level entity type has child entity types, though some level 2 entity types do not have level 3 child types.

It should be noted that each list of child entity types (level 2 and level 3) under a given parent is non-exhaustive, meaning there are sister level 2 or sister level 3 entity types that our current ontology does not include.

That does not mean we only annotate mentions referring to entities that fall into one of the terminal entity types. For example, the US vice president is an elected official, but there is no such entity type under PER.Politician. In this case, we should select PER.Politician as the entity type. Likewise, if a news article talks about immigration and mentions an immigrant called Jane Doe without any further information about her, we simply label Jane Doe as PER.

In the ontology spreadsheet, you will notice that for each parent type, there is an empty cell after that parent at the top of its child type column. That empty cell should be interpreted as "everything else" under that parent type. This is reflected in the in the annotation tool by an empty line at the top of every child entity type drop down list. So, if you do not make a choice from the list and leave it blank instead, the parent type is applied. In fact, any time you are in doubt, you should "back off" to the higher-level, more general classification.

The only exception is SituationalRole under PER. SituationalRole is a special entity type. All the entity types under it are specific to a situation or event. The 6 entity types, Patient, Protester, Survivor, Suspect, Victim and Voter are the only child types supported under SituationalRole. This is reflected by the missing empty cell after PER.SituationalRole in the spreadsheet. However, due to the design, the annotation tool still has a blank line on top of the 6 entity types. Be sure NOT to select this empty line.

3.4 Multiple Entity Types

A mention may be assigned more than one entity type:

- `President [Donald Trump] arrived at Walter Reed Medical Center in Bethesda, Maryland, Friday evening after experiencing a low-grade fever, chills, nasal congestion and cough.`

From the title "President", we know Donald Trump was the US president. We will label it as PER.Politician.HeadOfGovenment. He was taken to a hospital for treatment of COVID 19 infection. As it happens, we also have Patient as an entity type. Therefore, we must also assign PER.SituationalRole.Patient to the mention "Donald Trump".

Since we define the context as the paragraph in which the mention appears, it must be assigned ALL entity types as can be determined from the context.

- [Zelenskyy] grew up as a native Russian speaker in Kryvyi Rih, a major city of Dnipropetrovsk Oblast in central Ukraine. Prior to [his] acting career, [he] obtained a degree in law from the Kyiv National Economic University. [He] then pursued comedy and created the production company Kvartal 95, which produced films, cartoons, and TV shows including the TV series *Servant of the People*, in which [Zelenskyy] played the role of the Ukrainian president.

The mention of "Zelenskyy" in the first sentence would only be assigned "PER" if it were not for subsequent sentences, from which we learn that he studied at Kyiv National Economic University, pursued an acting career and founded the company Kvartal 95. We will therefore assign PER.Student, PER.Artist and PER.Executive.BusinessExecutive, not only to the mention of "Zelenskyy" in the first sentence, but also to mentions of the same entity everywhere else as indicated by pairs of brackets.

Another scenario where an entity mention may be assigned more than one entity type is when a pronoun refers to a conjoined antecedent with two (or more) mentions that refer to entities of different types. We had such an example in our sample document.

- The ACLU report lists examples of the FBI conducting surveillance on [protesters]$_{PER}$ and religious [groups]$_{ORG}$ with "aggressive tactics that infringe on [their]$_{\mathbf{PER;ORG}}$ free speech, religion and associational rights."

Here the pronoun *their*, via the mentions of *protesters* and *groups*, refers to entities of both PER and ORG. What entity type should we assign to the pronominal mention? For simplicity, the set of entity types for the pronominal mention with a conjoined antecedent is the union of all entity types applicable to the conjoined mentions.

If two entities share a higher-level entity type but we cannot assign a lower-level type for one of them, we will choose the higher-level type.

- President Joe Biden$_{\mathbf{PER.Politician.HeadOfGovernment}}$ and First Lady Jill Biden$_{\mathbf{PER}}$… They$_{\mathbf{PER}}$…

If we assign two entity types - PER.Politician.HeadOfGovernment and PER to "they", the end result is the same as assigning it with just PER.Politician.HeadOfGovernment. This is because the raw output expands a lower-level entity type to include its parent type (and grandparent type if applicable). Postprocessing removes duplicates. So assigning PER.Politician.HeadOfGovernment and PER to "they" will make it as if both Joe Biden and Jill Biden were heads of a government.

<u>New in 2022</u>:

We added Equipment as a level 1 entity type. This is a very broad entity type high in the Wikidata hierarchy. In the RUFES ontology, we have VEH(vehicle) and WEP(weapon) at the top level. In addition, we also ConsumerGoods.HomeAppliance and ConsumerGoods.MobileDevice. By Wikidata, these entity types (and their subtypes) are subtypes of Equipment. Since they can be automatically assigned (if desired), we will not label them as Equipment.

Note there is also MilitaryEquipment under Equipment. This is intended for NON-weapon military equipment.

Likewise we will not label an instance of PER as Animal. As an entity type in RUFES, PER is viewed as a social being.

4. Coreference

Coreference is a relationship between mentions – if two or more mentions in an article refer to the same entity, they corefer, for example, "James Comey" and "the new FBI director", "the ACLU" and "the American Civil Liberties Union", etc.

So, the third task of RUFES is to tag coreference within a document. This is done by associating all the mentions of that entity to an "entity name" (formally known an "entity ID") that must be unique for that entity in the document being annotated. There is a cell on the tool where you can either select an existing "entity name" that you have already entered or enter a new one. We must do so for every mention we select, or the tool will not allow us to continue.

4.1 Choice of Entity Name

Typically, you should use a named mention as the entity name wherever possible (the annotation tool has a copy button to let you paste the mention form over). For personal names, you would use the full name as the entity name. For longer organization names, an acronym or an abbreviation would be a better choice (the viewable cell is at a fixed length and may not display a long name completely on the screen). For entities with nominal mentions only, you may want to add a descriptive word to the mention head, especially if the news article contains several similar entities referred to by the same noun head but with different modifiers or in different paragraphs. Even for entities with named mentions, you may run into a situation where the same form is used for different entities. For example, if the news article mentions McDonald's as restaurants in different locations, then it is better to use *McDonald's at Tyson*, *McDonalds in Silver Spring,* etc. since they refer to different FAC entities. If McDonald's also refers to the corporation in another context, use *McDonalds_Company* as the company's entity name. These are just some general principles, although you can choose whatever form you like so long as it is easy for you and other annotators to recognize.

You can change an entity name that you have already entered, but the tool does not allow you to delete an existing entity name. While unused entity names have no effect on annotation quality, it can be annoying when the list has grown too long. Also, if you do change an entity name, you must go back to re-associate it with other mentions of the entity. Do not change an entity name simply because there is an extra space. You may be very confused if for example, "Barack Obama" and "Barack Obama" are both in the list. So, try not to make too many changes.

Never use a pronoun for an entity name. You should look for an antecedent or postcedent. Or in the case of a pronoun making a generic reference, e.g. "You shall not steal", use "people generic" as the entity name.

In the event when a reporter puts themselves in the news article by using the first person pronoun "I" (or "we") but nowhere in the article is the reporter's name mentioned, use something like "article reporter".

4.2 Generic and Indefinite Reference

Some entity tasks have started labeling generic reference. While RUFES currently does not have a separate label to distinguish between generic and specific references, we will be indicating generic reference by putting the word "generic" at the end of the entity name. For example,

- John[John] is a good linguist[good linguist generic].

As noted before, "John" and "a good linguist" do not refer to the same entity. But,

- John[John] is the good linguist[John].

Specific and generic references should never corefer – a specific reference is an instance of a generic (type) reference and they never refer to the same entity.

We mention earlier that "false" titles are treated similarly as appositives with the following examples:

- cellist Yo-Yo Ma
- famed New Left philosopher Herbert Marcuse
- convicted bomber Timothy McVeigh

How do we determine if the reference of a false title? It may be helpful to covert the construction into apposition:

- Yo-Yo Ma, a cellist
- Herbert Marcuse, the famed New Left philosopher
- Timothy McVeigh, a/the convicted bomber

In the first example, "a cellist" should be a generic reference. In the second and third examples, the false tiles are more likely specific references.

What do you think of "a niece" below?

- Mary L. Trump, a niece of the president who wrote a scathing bestseller documenting deep family dysfunction, has sued President Trump and two of his siblings.

Be careful about indefinite references. Two mentions that look the same or similar may not refer to the same set of entities. So, to distinguish between different entities, some descriptive words should be added to the entity name. For example,

- Some linguists like potato, some linguists like tomato.

We may use "potato linguists" and "tomato linguists" for the two sets of linguists.

How should we annotate the following paragraph?

> A journalist questioned Dr. Jay Butler, deputy director for infectious diseases at the Centers for Disease Control and Prevention (CDC) and Dr. Andrew Pavia, pediatric infectious diseases specialist at University of Utah School of Medicine, over potential safety concerns behind a third dose, noting some fully vaccinated people have sought out booster shots on their own.

4.3 Mentions in Article Title

Titles of a news article tend to be short, often omitting function words. So, the context of a title may be limited for certain mentions, especially nominal mentions. It is recommended you do not immediately tag mentions in a title. In fact, the best practice is to quickly go over the entire article so that you'll have a basic idea on what entities are expected. You may wait until you finish the entire article before annotating the title.

4.4 Coreference and Conjoined Mentions

If you recall, we do not add a third mention to a conjunction of two (or more) mentions. What if a later mention refers to the conjunction? Here is an example:

- `[US] and [China] have been engaged in a trade war. The two [countries] have raised tariffs on [each other]'s exports.`

We will tag "US" and "China" as two separate mentions. We do not create a third mention of [US and China] to refer to both countries. But the nominal and pronominal mentions in the second sentence of the paragraph clearly refer to "US and China". In this case, we will create an entity name such as "US & China" and apply it to both "the two countries" and "each other".

4.5 Mentions in Negative Context

As mentioned in the PRO section, negative mentions and mentions in a negative context are potentially taggable – another departure from RUFES 2020. But how do we do coreference? Consider again the two examples.

- `Although Jenkins and Rankovicz did well in the primaries, neither won a majority of the vote.`
- `Although Jenkins and Rankovicz did well in the primaries, both failed to win a majority of the vote.`

We will "move" negation to the predicate and treat the negative mention as "positive". So, we can use "Jenkins & Rankovicz" for both "neither" and "both".

Mentions in a negative context may be generic as shown in an earlier example, repeated below:

- `The US and its coalition did not find [weapons] of mass destruction in Iraq.`

Similarly,

- `Homebuyers are hunting for houses in more rural areas, but no one is interested in McMansions anymore.`

We consider "no one" refers to the same entity as "homebuyers" and may use "homebuyers generic" as the entity name.

5. Complete List of Entity Types, Definitions and Examples

(Refer to the ontology spreadsheet)

6. Chinese Annotation Supplement

While overall annotation guidelines (such as entity types, mention types, co-reference, etc.) are not language-specific, languages may differ on how they may be applied due to each language's linguistic characteristics. This section highlights some rules and principles that are specific to Chinese. Annotators are encouraged to consult with the team lead and/or with each other for difficult decisions.

6.1 Wordhood and Mention Extent

Regardless how it is defined, the notation of word in Chinese is not so straightforward to native speakers not trained in linguistics. But even linguists do not always agree on how it is defined and applied to Chinese. This is because Chinese morphemes are largely monosyllabic, represented by single characters in writing. There are a small number of multisyllabic morphemes (mostly disyllabic whereas morphemes with 3 or more syllables are usually loan words), but even so, each syllable in a multisyllabic morpheme is still represented by one character, and except for some loan words (e.g., 琵琶 "pipa"), it may be used as a morpheme on its own (e.g., 蝴蝶 "butterfly", where 蝶 itself also means "butterfly").

Chinese has not been actively "inventing" new monosyllabic morphemes/characters. As a result, most "words" in modern Chinese vocabulary are formed by combining existing characters. Compounding is the predominant morphological process. Affixation is also productive, though the distinction between affixes (bound morphemes) and free morphemes is not always obvious.

6.1.1 Nominal Mention

Compounds

In English, a distinction is made between open and close compounds, and for an open compound, only the head of the compound is selected to represent the mention. English also treats hyphenated compounds as close compounds. While there is no orthographic clue for this kind of differentiation in Chinese, the word length plays an important: the shorter a compound is, the more likely it is a close one. In general, a compound with 2 and 3 characters tend to a close compound, which is treated as the head of a mention in RUFES.

- 汽车(car), 火车(train), 轮船(propeller ship)
- 口罩(face mask), 手套 (glove)
- 冰箱(refrigerator), 电话(telephone), 电视机(television), 电脑(computer), 手机(cell phone)
- 餐馆 (restaurant), 饭店 (restaurant/hotel)
- 炮弹(shell), 战(斗)机(fighter jet)
- 网页(webpage), 网站(website)
- 白人(white person), 黑人(black person)
- 候机楼 (airport terminal), 旅行社 (travel agency)
- 左派 (leftist), 反对派 (opposition side)

A compound with 4 or more characters is more likely an open compound:

- 高速[公路] (highway), 飞机[跑道] (runway)
- 空气[压缩机] (air compressor)

As shown above, the head of an open compound can be used to refer to the same entity. If the head is not usually used to refer to the same entity, the entire compound should be treated as the mention head. For example, while 航空母舰 (aircraft carrier) is a type of "mother ship", it is not common to replace it with 母舰 (which literally means "mother ship") in the same context. Instead, its abbreviated form 航母 is often used. Therefore, 航空母舰 should not be further segmented.

Affixation

The number of affixes in Chinese is quite small. As applied to nouns, the following is a list of most commonly used suffixes:

- 子 (forming a disyllabic word): 孩子 (child), 房子 (house), 猴子 (monkey)
- 头 (also forming a disyllabic word): 馒头 (bun), 骨头 (bone)
- 儿 (diminutive suffix often used in northern dialects): 门儿 (door), 船儿 (boat)
- 们 (plural suffix for people): 同志们 (comrades), 朋友们 (friends)
- 者 (used with a verb or an adjective to form a noun): 学者 (scholar), 作者 (author), 伤者 (the injured)
- 家 (specialist): 数学家 (mathematician), 物理学家 (physicist), 作家 (writer)
- 员 (member): 演员 (actor), 党员 (party member), 会员 (association member),程序员 (programmer)

Suffixes as shown above must be included with the stem which they are affixed with and the whole word should be marked as the head of a nominal mention.

Resulting words from affixation appearing as part of a compound are subject to the same principles on compounds:

- 先进[工作者] (model worker), 体力[劳动者] (laborer), 科技[工作者] (science worker)

Affixation, like compounding, can be recursive. But unlike recursive compounding, recursive affixation can result in long close compounds that must be treated as mention heads:

- [无政府主义者] (anarchist), [共产主义者] (communist)

主义 (-ism) is a disyllabic suffix reborrowed from Japanese to form words referring to the kind of ideology expressed by the stem. So, 共产主义 and 无政府主义 should be treated as words. New words are formed by adding the suffix 者 to those words and they should not be segmented either, as shown in the above two examples. Note how they are different from the following open compound:

- 共产主义[信仰者] (communism believer)

Words such as 人员 (personnel) and 分子 (a person of a social group) are not suffixes. The following examples therefore should be treated as open compounds.

- 医护[人员] (health personnel)，恐怖[分子] (terrorist)

Abbreviation

Chinese tend to shorten long compounds to 2 or 3 characters. In general, these abbreviations (consisting of one character from two or more component words of a compound) should not be further segmented.

- [航空母舰] -> [航母] (aircraft carrier)
- 空气[压缩机] -> [空压机] (air compressor)

Loan Words

Loan words are not to be segmented even if part of the word is a native morpheme:

- 巧克力(chocolate), 雪茄(cigar), 沙发(sofa)
- 汉堡包(hamburger), 路由器(router)

的 as the nominal head

The auxiliary particle 的 can be used before a noun to show possessiveness or descriptive attribute ascribed to the proceeding noun, pronoun, adjective or clause. Given the right context, the noun after 的 may be omitted.

- 张三的孩子上北大，李四[的]上清华 (Zhang San's son attends Peking University. Li Si' attends Tsinghua University)
- 张三买了一辆红色的车，李四买了一辆白色[的] (Zhang San bought a red car. Li Si bought a white one)

In such cases, we will label 的 as the head of a nominal mention as shown above.

6.1.2 Named Mention

Since the extent of a named mention is the entire proper noun phrase and Chinese does not even have articles such as "the", marking the boundary of a named mention in Chinese is relatively easy.

- 北京(Beijing), 北京市(Beijing City), 南京(Nanjing), 南京市(Nanjing City), 广东省(Guangdong Province)
- 华为技术有限公司(Huawei Technologies Co., Ltd.), 苹果公司 (Apple Inc.), 达美航空 (Delta Air Lines)

Note an entity may choose a name in a foreign language that deviates from the one in the original language. A notable example is United Airlines, which is commonly known as 美联航. Here 美 (America) should not be tagged separately to refer to "United States". But 美国谷歌 (America's Google) should be tagged as two entity mentions.

Affixes used with personal names are included in the mention extent:

- 老张, 小王, 阿宝
- 冬儿, 祥子

When a phrase containing two (or more) nested named mentions referring to two entities, one being part of/subordinate to the other, we tag them as separate entity mentions.

- [中国][外交部] (Chinese Foreign Ministry), [江苏省][南京市] (Jiangsu Nanjing)

Do not tag a separate entity mention if there is no such relationship.

- 北京大学 (Peking University) -> [北京大学] not [北京][大学]

When a phrase with multiple named mentions refer to two or more entities, their abbreviated forms (often a single character from each proper noun) may be used, e.g., 中美关系 (Sino-US relationship), 中日韩首脑 (leaders from China, Japan, and South Korea). In such cases, each abbreviation should be tagged on its own, hence [中][美]关系.

6.1.3 Special Cases

Compounding is not limited to stems that are common nouns. Proper nouns can participate in compounding too.

In English, the morphology of forming a person's locality origin is quite complicated. Affixation is one of the most common methods, e.g., adding suffixes such as -ese, -(a)n, -er, etc.

- America -> American
- Japan -> Japanese
- New York -> New Yorker
- Canada -> Canadian
- Mexico -> Mexican

Sometimes the word formation is backwards or completely "irregular":

- Turkey <- Turk
- Greece -> Greek

Regardless, all these nouns referring to persons from a locality must be treated as one mention since they are at the word level and cannot be further segmented.

However, the process in Chinese is very simple – by joining the character 人 (person) to the end of the name of a locality, the expression referring a person from that locality is formed.

- 美国->美国人 (American)
- 英国->英国人 (English/Brit)
- 中国->中国人 (Chinese)
- 俄罗斯->俄罗斯人 (Russian)
- 新加坡->新加坡人 (Singaporean)
- 弗吉尼亚->弗吉尼亚人 (Virginian)
- 上海->上海人 (Shanghainese)
- 北京->北京人 (Beijinger)

These compounds feel more like open than close compounds to native speakers. For RUFES, they will be treated as two mentions referring to two entities, a GPE/LOC entity and a PER entity. Like so, [美国][人], [北京][人].

However, when 人 is used with an ethnicity name, we will not mark two mentions (like one referring to the ethnic group and one to an individual from the group). So, 汉族人 (Han Chinese), 维吾尔(族)人 (Uyghur), (西)藏人 (Tibetan), 华人 (Chinese), 拉美人 (Latino), etc. will only be tagged as one PER mention.

Some foreign currency names in Chinese consist of a country name (often abbreviated) followed by the common noun 元, e.g., 美元 (American dollar), 加(拿大)元 (Canadian dollar), 欧元 (Euro), etc. But many other currency names are transliterated, for example, 法郎 (franc), 卢布 (Rubi), 卢比 (rupee), and country names are prefixed to the transliteration to refer to country-specific currencies such as 法国法郎 (French franc), 中非法郎 (Central African franc), 俄罗斯卢布 (Russian Rubi), 巴基斯坦卢比 (Pakistani rupee), etc. It is just the word dollar is translated into Chinese as 元, a currency unit native in China. For RUFES, we consider currency names "atomic" words and thus do not treat the country name as a separate entity mention.

6.2 Entity Typing

Entity types are language independent. However due to the uniqueness of China's political and economic systems, annotators may have difficulty classifying certain entities specific to China, particularly for government and party related organizations.

We limit the definition of a government organization to those directly associated with a GPE entity that performs governing duties for the GPE entity. Although most hospitals, schools, research institutes and even enterprises are state-owned in China, they will not be treated as government organizations and should be labeled as commercial, health, educational, etc. according to their purported function.

There is another complexity that is unique to a communist country, the omnipresence of the communist party at every level of government as well as in military and state-owned organizations (and even in some private organizations too). In China for example, at the highest level there is 中国共产党中央委员会 (Central Committee of the Chinese Communist Party), then 省委 (provincial committee), etc. until at the lowest level 村支部 (village branch). These organizations exert more power than their corresponding administrative organizations because they are responsible for policy making and personnel decisions, among many other important duties. But they are not directly involved in daily operation of the government. For RUFES, they will be labeled as ORG.PoliticalOrganization whereas their corresponding administrative organizations will be assigned ORG.Government. This also applies to state-owned educational, financial, commercial, and other non-governing organizations.

This difference in power in also reflected in personnel for government and state-owned organizations. For government organizations, especially at and above the county-level, the party secretary and the governor/mayor are usually two different people, and the government/mayor is only appointed as depute secretary. But since government officials are all considered politicians, the distinction does not affect labeling for leaders at various levels of GPE. For non-governing organizations, the top party leader and the highest-level executive are often the same person. For example, 马永生 is both Chairman (of the board) and Secretary (of the party member group) at Sinopec Group. 赵东 is President but Deputy Secretary. We continue to apply PER.Politician to mentions referring to positions in the party. Thus, if both positions (party and operational) are mentioned, the entity will be assigned two entity types. Private sectors do not usually have party groups, but if a party group and its leader are mentioned, they

will be labeled as ORG.PoliticalOrganization and PER.Politician, respectively. If there is no specific reference to party leadership for a person entity of leadership, the person should be labeled as PER.Executive (and one of its subtypes as determined by the context).

Appendix I: Sample Document

ACLU calls on Obama, Congress to rein in power of the FBI

Monday, September 16, 2013 20:01:18 GMT-05:00

As James Comey takes over as the new FBI director, the American Civil Liberties Union is calling on the Obama administration and Congress to rein in the increasing power of the agency.

In a critical 63-page report that will be issued Tuesday, the ACLU says the powers of the FBI have expanded too dramatically over the past 12 years, transforming the Bureau into a "secret domestic intelligence agency."

"The excessive secrecy with which it cloaks these domestic intelligence gathering operations has crippled constitutional oversight mechanisms," the report says. "Courts have been reticent to challenge government secrecy demands and, despite years of debate in Congress . . . it took unauthorized leaks by a whistleblower to finally reveal the government's secret interpretation of these laws and the Orwellian scope of its domestic surveillance programs."

The ACLU report, entitled "Unleashed and Unaccountable: The FBI's Unchecked Abuse of Authority," compiles examples of the changes of law and policy since the Sept. 11, 2001, attacks on the United States, which the group says "unleashed the FBI from its traditional restraints and opened the door to abuse."

A spokesman for the FBI said he could not comment on the report because FBI officials had not yet seen it.

The changes highlighted in the report include the FBI's racial and ethnic mapping program, which allows the FBI to collect demographic information to map American communities by race and ethnicity; the use of secret National Security Letters, which asked for account information from telecommunications companies, financial institutions and credit agencies and required no judicial approval; warrantless wiretapping; and the recent revelations about the government's use of Section 215 of the Patriot Act to track all U.S. telephone calls.

In its report, the ACLU asks Congress, the president and the attorney general to conduct a comprehensive evaluation of the FBI's policies and programs, and makes 15 recommendations for reform of the agency.

"The list of abuses is long and demonstrates that Congress must do a top-to-bottom review of FBI politics and practices to identify and curtail any activities that are unconstitutional or easily

misused," said Hina Shamsi, director of the ACLU's National Security Project. "The time for wholesale reform has come."

The ACLU report lists examples of the FBI conducting surveillance on protesters and religious groups with "aggressive tactics that infringe on their free speech, religion and associational rights."

The FBI's increased intelligence collection powers have led to a data explosion that agents cannot keep up with, making it harder for the agency to focus on suspects and groups that should be investigated, the ACLU concluded in its report.

"Rather than aiding its terrorism prevention efforts, the FBI's expanded investigative and intelligence powers have overwhelmed agents with a flood of irrelevant information and false alarms," said Michael German, senior policy counsel at the ACLU's Washington Legislative Office and a former FBI agent.

As an example, the report cites the unanswered questions surrounding the FBI's three-month investigation of Tamerlan Tsarnaev, one of the Boston Marathon bombing suspects before the deadly attacks.

"FBI agents cannot be expected to be fortune tellers," the report said. "But reviewing the facts of this matter is important to determine whether current FBI practices are effective. Its investigation of Tsarnaev was one of over 1,000 assessments the Boston Joint Terrorism Task Force completed in 2011. . . . This torrid pace may have diminished the quality of the Tsarnaev assessment."

Appendix II: Sample Document Annotation Output (Selected Fields)

Mention String	Offsets	Entity ID	Entity Types	Mention Type
ACLU	0-3	ACLU	ORG;ORG.AdvocacyGroup	NAM
Obama	14-18	B Obama	PER;PER.Politician;PER.Politician.HeadOfGovernment	NAM
Congress	21-28	US Congress	ORG;ORG.Government;ORG.Government.Legislature	NAM
FBI	54-56	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
James Comey	107-117	James Comey	PER;PER.CivilServant	NAM
FBI	141-143	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
director	145-152	James Comey	PER;PER.CivilServant	NOM
American Civil Liberties Union	159-188	ACLU	ORG;ORG.AdvocacyGroup	NAM
Obama	208-212	B Obama	PER;PER.Politician;PER.Politician.HeadOfGovernment	NAM

administration	214-227	US executive branch	ORG;ORG.Government;ORG.Government.Administration	NOM
Congress	233-240	US Congress	ORG;ORG.Government;ORG.Government.Legislature	NAM
agency	281-286	FBI	ORG;ORG.Government;ORG.Government.Agency	NOM
report	311-316	ACLU report	Document;Document.Report	NOM
that	318-321	ACLU report	Document;Document.Report	PRO
ACLU	351-354	ACLU	ORG;ORG.AdvocacyGroup	NAM
FBI	379-381	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
Bureau	455-460	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
agency	499-504	secret agency generic	ORG;ORG.Government;ORG.Government.Agency	NOM
it	542-543	FBI	ORG;ORG.Government;ORG.Government.Agency	PRO
report	656-661	ACLU report	Document;Document.Report	NOM
Courts	670-675	courts generic	ORG;ORG.Court	NOM
Congress	768-775	US Congress	ORG;ORG.Government;ORG.Government.Legislature	NAM
whistleblower	815-827	a whistleblower	PER	NOM
government	851-860	US government	ORG;ORG.Government	NOM
laws	895-898	oversight laws	LAW	NOM
its	927-929	US government	ORG;ORG.Government	PRO
ACLU	968-971	ACLU	ORG;ORG.AdvocacyGroup	NAM
report	973-978	ACLU report	Document;Document.Report	NOM
Unleashed and Unaccountable : The FBI's Unchecked Abuse of Authority	991-1057	ACLU report	Document;Document.Report	NAM
law	1097-1099	US laws generic	LAW	NOM
United States	1153-1165	USA	GPE;GPE.Country	NAM
group	1178-1182	ACLU	ORG;ORG.AdvocacyGroup	NOM
FBI	1204-1206	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM

its	1213-1215	FBI	ORG;ORG.Government;ORG.Government.Agency	PRO
spokesman	1273-1281	FBI spokesman	PER;PER.CivilServant;PER.CivilServant.Spokesperson	NOM
FBI	1291-1293	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
he	1300-1301	FBI spokesman	PER;PER.CivilServant;PER.CivilServant.Spokesperson	PRO
report	1328-1333	ACLU report	Document;Document.Report	NOM
FBI	1343-1345	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
officials	1347-1355	FBI	ORG;ORG.Government;ORG.Government.Agency	NOM
it	1374-1375	ACLU report	Document;Document.Report	PRO
report	1409-1414	ACLU report	Document;Document.Report	NOM
FBI	1428-1430	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
FBI	1486-1488	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
American	1532-1539	USA	GPE;GPE.Country	NAM
communities	1541-1551	American communities generic	PER	NOM
National Security Letters	1594-1618	National Security Letters	Document;Document.LegalDocument	NAM
which	1621-1625	National Security Letters	Document;Document.LegalDocument	PRO
companies	1681-1689	tele companies generic	ORG;ORG.CommercialOrganization;ORG.CommercialOrganization.UtilityCompany	NOM
institutions	1702-1713	fin institutions generic	ORG;ORG.CommercialOrganization;ORG.CommercialOrganization.FinancialInstitution	NOM
agencies	1726-1733	credit agencies generic	ORG;ORG.CommercialOrganization;ORG.CommercialOrganization.FinancialInstitution	NOM
government	1832-1841	US government	ORG;ORG.Government	NAM
Section 215	1852-1862	Section 215	LAW	NAM
Patriot Act	1871-1881	Patriot Act	LAW	NAM
U.S	1896-1898	USA	GPE;GPE.Country	NAM
its	1921-1923	ACLU	ORG;ORG.AdvocacyGroup	PRO
ACLU	1937-1940	ACLU	ORG;ORG.AdvocacyGroup	NAM
Congress	1947-1954	US Congress	ORG;ORG.Government;ORG.Government.Legislature	NAM

president	1961-1969	B Obama	PER;PER.Politician;PER.Politician.HeadOfGovernment	NOM
attorney	1979-1986	attorney general	PER;PER.CivilServant;PER.CivilServant.Minister	NOM
FBI	2041-2043	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
agency	2117-2122	FBI	ORG;ORG.Government;ORG.Government.Agency	NOM
Congress	2175-2182	US Congress	ORG;ORG.Government;ORG.Government.Legislature	NAM
FBI	2218-2220	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
Hina Shamsi	2335-2345	Hina Shamsi	PER;PER.Executive	NAM
director	2348-2355	Hina Shamsi	PER;PER.Executive	NOM
ACLU	2364-2367	ACLU	ORG;ORG.AdvocacyGroup	NAM
National Security Project	2371-2395	ACLU Nat Sec Proj	ORG;ORG.AdvocacyGroup	NAM
ACLU	2444-2447	ACLU	ORG;ORG.AdvocacyGroup	NAM
report	2449-2454	ACLU report	Document;Document.Report	NOM
FBI	2478-2480	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
protesters	2509-2518	protestors generic	PER;PER.SituationalRole.Protester	NOM
groups	2534-2539	religious groups generic	ORG;ORG.ReligiousOrganization	NOM
their	2583-2587	protestors and religious groups generic	PER;PER.SituationalRole.Protester;ORG;ORG.ReligiousOrganization	PRO
FBI	2642-2644	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
agents	2723-2728	FBI agents generic	PER;PER.CivilServant	NOM
agency	2776-2781	FBI	ORG;ORG.Government;ORG.Government.Agency	NOM
suspects	2795-2802	key suspects generic	PER;PER.SituationalRole.Suspect	NOM
groups	2808-2813	key groups generic	ORG	NOM
ACLU	2848-2851	ACLU	ORG;ORG.AdvocacyGroup	NAM
its	2866-2868	ACLU	ORG;ORG.AdvocacyGroup	PRO
report	2870-2875	ACLU report	Document;Document.Report	NOM
its	2898-2900	FBI	ORG;ORG.Government;ORG.Government.Agency	PRO

FBI	2936-2938	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
agents	3006-3011	FBI agents generic	PER;PER.CivilServant	NAM
Michael German	3076-3089	Michael German	PER;PER.Professional;PER.Professional.Lawyer;PER.CivilServant	NAM
counsel	3106-3112	senior counselors generic	PER;PER.Professional	NAM
ACLU	3121-3124	ACLU	ORG;ORG.AdvocacyGroup	NAM
Washington Legislative Office	3128-3156	ACLU's DC Leg Off	ORG;ORG.AdvocacyGroup	NAM
FBI	3171-3173	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
agent	3175-3179	former FBI agents generic	PER;PER.CivilServant	NOM
report	3201-3206	ACLU report	Document;Document.Report	NOM
FBI	3255-3257	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
Tamerlan Tsarnaev	3290-3306	Tamerlan Tsarnaev	PER;PER.SituationalRole.Suspect	NAM
one	3309-3311	Tamerlan Tsarnaev	PER;PER.SituationalRole.Suspect	NOM
Boston	3320-3325	city of Boston	GPE;GPE.City	NAM
suspects	3344-3351	Boston bombing suspects	PER;PER.SituationalRole.Suspect	NOM
FBI	3381-3383	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
agents	3385-3390	FBI agents generic	PER;PER.CivilServant	NOM
tellers	3425-3431	fortune tellers generic	PER	NOM
report	3439-3444	ACLU report	Document;Document.Report	NOM
FBI	3534-3536	FBI	ORG;ORG.Government;ORG.Government.Agency	NAM
Its	3563-3565	FBI	ORG;ORG.Government;ORG.Government.Agency	PRO
Tsarnaev	3584-3591	Tamerlan Tsarnaev	PER;PER.SituationalRole.Suspect	NAM
Boston Joint Terrorism Task Force	3631-3663	Boston Joint Terr TF	ORG;ORG.Government	NAM

Tsarnaev	3746-3753	Tamerlan Tsarnaev	PER;PER.SituationalRole.Suspect	NAM

Appendix III: Entity Type Cheat Sheet

Level 1	**Level 2**	**Level 3**
Animal		
Animal	DomesticatedAnimal	
Animal	DomesticatedAnimal	Livestock
Animal	DomesticatedAnimal	Pet
APP		
APP	CommunicationSoftware	
APP	CommunicationSoftware	SocialMedia
AstroObject		
AstroObject	Comet	
AstroObject	Galaxy	
AstroObject	NaturalSatellite	
AstroObject	Planet	
AstroObject	Star	
ConsumerGoods		
ConsumerGoods	CleaningProduct	
ConsumerGoods	CleaningProduct	Detergent
ConsumerGoods	CleaningProduct	Soap
ConsumerGoods	Clothing	
ConsumerGoods	DisinfectingProduct	
ConsumerGoods	Food	
ConsumerGoods	Furniture	
ConsumerGoods	Furniture	SeatingFurniture
ConsumerGoods	Furniture	SleepingFurniture
ConsumerGoods	Furniture	StorageFurniture
ConsumerGoods	Furniture	Table
ConsumerGoods	HomeAppliance	
ConsumerGoods	HomeAppliance	MajorAppliance
ConsumerGoods	HomeAppliance	SmallAppliance
ConsumerGoods	MobileDevice	
ConsumerGoods	MobileDevice	Smartphone
ConsumerGoods	MobileDevice	TabletComputer
ConsumerGoods	PPE	
ConsumerGoods	PPE	FaceShield
ConsumerGoods	PPE	Glove
ConsumerGoods	PPE	ProtectiveClothing
ConsumerGoods	PPE	ProtectiveEyewear
ConsumerGoods	PPE	ProtectiveFaceMask

ConsumerGoods	RecreationalDrug	
Document		
Document	LegalDocument	
Document	LegalDocument	Ballot
Document	LegalDocument	Certificate
Document	LegalDocument	PersonalIdentification
Document	Letter	
Document	Letter	Email
Document	Map	
Document	Map	DigitalMap
Document	Report	
Document	TextMessage	
Equipment		
Equipment	AgriculturalEquipment	
Equipment	IndustrialEquipment	
Equipment	MedicalEquipment	
Equipment	MedicalEquipment	DiagnosticEquipment
Equipment	MedicalEquipment	LifeSupportEquipment
Equipment	MedicalEquipment	MedicalLabEquipment
Equipment	MedicalEquipment	MedicalMonitor
Equipment	MedicalEquipment	TherapeuticEquipment
Equipment	MilitaryEquipment	
Equipment	SportsEquipment	
FAC		
FAC	Building	
FAC	Building	ApartmentBuilding
FAC	Building	Clinic
FAC	Building	Courthouse
FAC	Building	GovernmentBuilding
FAC	Building	Hospital
FAC	Building	House
FAC	Building	Jail
FAC	Building	OfficeBuilding
FAC	Building	PlaceOfWorship
FAC	Building	PoliceStation
FAC	Building	School
FAC	Building	StoreShop
FAC	Building	Warehouse
FAC	Depository	
FAC	Depository	Bank
FAC	Depository	Library
FAC	Depository	Museum
FAC	GardenPark	

FAC	GardenPark	Garden
FAC	GardenPark	Park
FAC	GardenPark	Zoo
FAC	MilitaryInstallation	
FAC	PollingPlace	
FAC	Structure	
FAC	Structure	Barricade
FAC	Structure	Bridge
FAC	Structure	Monument
FAC	Structure	Plaza
FAC	Structure	Tower
FAC	TransportHub	
FAC	TransportHub	Airport
FAC	TransportHub	BusStation
FAC	TransportHub	Port
FAC	TransportHub	TrainStation
FAC	Way	
FAC	Way	Canal
FAC	Way	Highway
FAC	Way	RailroadTrack
FAC	Way	Street
FAC	Way	Tunnel
GPE		
GPE	City	
GPE	Country	
GPE	County	
GPE	ElectoralDistrict	
GPE	NonSovereignCountry	
GPE	ProvinceState	
GPE	Territory	
GPE	Town	
GPE	UnionOfCountries	
GPE	Village	
IllHealth		
IllHealth	Disease	
IllHealth	Disease	CommunicableDisease
IllHealth	Disease	NonCommunicableDisease
IllHealth	Injury	
IllHealth	SignOrSymptom	
LAW		
LAW	Bill	
LAW	Referendum	
LAW	Treaty	

LOC		
LOC	AirSpace	
LOC	Border	
LOC	Border	PoliticalBorder
LOC	CrimeScene	
LOC	GeographicPoint	
LOC	GeographicPoint	Address
LOC	GeographicPoint	CheckPoint
LOC	Land	
LOC	Land	Archipelago
LOC	Land	Continent
LOC	Land	Field
LOC	Land	Island
LOC	Land	Mountain
LOC	Land	MountainRange
LOC	Land	Peninsula
LOC	Mine	
LOC	Neighborhood	
LOC	Region	
LOC	Water	
LOC	Water	Lake
LOC	Water	Ocean
LOC	Water	River
LOC	Water	Sea
MED		
MED	Vaccine	
MED	Vaccine	AttenuatedVaccine
MED	Vaccine	InactivatedVaccine
MED	Vaccine	RnaVaccine
MED	Vaccine	SubunitVaccine
MED	Vaccine	VirVecVaccine
MED	Antimicrobial	
MED	Painkiller	
MaterialAnatomicalEntity		
MaterialAnatomicalEntity	AnatomicalStructure	
MaterialAnatomicalEntity	AnatomicalStructure	CardinalBodyPart
MaterialAnatomicalEntity	AnatomicalStructure	Cell
MaterialAnatomicalEntity	AnatomicalStructure	Organ
MaterialAnatomicalEntity	AnatomicalStructure	OrganSystem
MaterialAnatomicalEntity	AnatomicalStructure	SubCardinalBodyPart
MaterialAnatomicalEntity	BiogenicSubstance	
MaterialAnatomicalEntity	BiogenicSubstance	BodyFluids
ORG		

ORG	AdvocacyGroup	
ORG	Association	
ORG	Association	Club
ORG	Association	GreekLife
ORG	Association	League
ORG	Association	ProfessionalAssociation
ORG	Association	TradeAssociation
ORG	Charity	
ORG	CommercialOrganization	
ORG	CommercialOrganization	BroadcastingCompany
ORG	CommercialOrganization	EnergyCompany
ORG	CommercialOrganization	FinancialInstitution
ORG	CommercialOrganization	Firm
ORG	CommercialOrganization	Manufacturer
ORG	CommercialOrganization	NewsAgency
ORG	CommercialOrganization	PharmaceuticalCompany
ORG	CommercialOrganization	Retailer
ORG	CommercialOrganization	TechCompany
ORG	CommercialOrganization	TransportCompany
ORG	CommercialOrganization	UtilityCompany
ORG	Court	
ORG	Court	InternationalCourt
ORG	Court	LocalCourt
ORG	Court	MilitaryCourt
ORG	Court	NationalCourt
ORG	Court	SupremeCourt
ORG	CriminalOrganization	
ORG	CriminalOrganization	Mafia
ORG	CriminalOrganization	StreetGang
ORG	CriminalOrganization	TerroristGroup
ORG	EducationalInstitution	
ORG	EducationalInstitution	College
ORG	EducationalInstitution	GradeSchool
ORG	EducationalInstitution	LanguageSchool
ORG	EducationalInstitution	Preschool
ORG	EducationalInstitution	SecondarySchool
ORG	EducationalInstitution	TrainingSchool
ORG	Government	
ORG	Government	Administration
ORG	Government	Agency
ORG	Government	ArmedForces
ORG	Government	FireDepartment
ORG	Government	LegislativeCommittee

ORG	Government	Legislature
ORG	Government	Ministry
ORG	Government	Police
ORG	NonGovernmentMilitary	
ORG	NonGovernmentMilitary	InsurgentOrganization
ORG	NonGovernmentMilitary	MercenaryOrganization
ORG	NonGovernmentMilitary	Militia
ORG	HealthcareInstitution	
ORG	PoliticalOrganization	
ORG	PoliticalOrganization	LaborUnion
ORG	PoliticalOrganization	PAC
ORG	PoliticalOrganization	Party
ORG	ReligiousOrganization	
ORG	Team	
ORG	Team	SportsTeam
ORG	WorldOrganization	
Pathogen		
Pathogen	Bacterium	
Pathogen	Virus	
Pathogen	Virus	Coronavirus
Pathogen	Virus	InfluenzaVirus
Pathogen	Virus	Rhinovirus
PER		
PER	Royal	
PER	Royal	Monarch
PER	Artist	
PER	Artist	Musician
PER	Artist	Painter
PER	Artist	Photographer
PER	CivilServant	
PER	CivilServant	Ambassador
PER	CivilServant	Judge
PER	CivilServant	Minister
PER	CivilServant	PoliceOfficer
PER	CivilServant	PolicyAdvisor
PER	Criminal	
PER	Executive	
PER	Executive	AcademicAdministrator
PER	Executive	BusinessExecutive
PER	Farmer	
PER	Laborer	
PER	Laborer	Farmworker
PER	Laborer	Miner

PER	Politician	
PER	Politician	Candidate
PER	Politician	Elected
PER	Politician	Governor
PER	Politician	HeadOfGovernment
PER	Politician	HeadOfNation
PER	Politician	Legislator
PER	Politician	Mayor
PER	Politician	PartyCommitteeSecretary
PER	Politician	PartyGeneralSecretary
PER	Professional	
PER	Professional	Athlete
PER	Professional	CampaignManager
PER	Professional	Firefighter
PER	Professional	HealthProfessional
PER	Professional	Journalist
PER	Professional	Lawyer
PER	Professional	Lobbyist
PER	Professional	Mercenary
PER	Professional	Professor
PER	Professional	SchoolTeacher
PER	Professional	Scientist
PER	Professional	Spokesperson
PER	ReligiousLeader	
PER	Serviceman	
PER	Serviceman	MilitaryOfficer
PER	Serviceman	Veteran
PER	SituationalRole	Patient
PER	SituationalRole	Protester
PER	SituationalRole	Survivor
PER	SituationalRole	Suspect
PER	SituationalRole	Victim
PER	SituationalRole	Voter
PER	Student	
PER	Terrorist	
PER	Writer	
PER	Writer	Author
PER	Writer	Speechwriter
Publication		
Publication	Book	
Publication	Magazine	
Publication	Magazine	NewsMagazine
Publication	Magazine	PopularMagazine

Publication	Magazine	ReligiousMagazine
Publication	Magazine	ScholarlyJournal
Publication	Magazine	TradeMagazine
Publication	Newspaper	
Publication	Post	
Publication	Post	BlogPost
Publication	Website	
VEH		
VEH	Aircraft	
VEH	Aircraft	CivilAircraft
VEH	Aircraft	Drone
VEH	Aircraft	MilitaryAircraft
VEH	Rocket	
VEH	Spacecraft	
VEH	Spacecraft	Satellite
VEH	Spacecraft	SpaceStation
VEH	Watercraft	
VEH	Watercraft	Boat
VEH	Watercraft	CargoShip
VEH	Watercraft	CruiseShip
VEH	Watercraft	Warship
VEH	Watercraft	Yacht
VEH	WheeledVehicle	
VEH	WheeledVehicle	ArmoredVehicle
VEH	WheeledVehicle	HumanPoweredVehicle
VEH	WheeledVehicle	MotorVehicle
VEH	WheeledVehicle	RailedVehicle
WEA		
WEA	CloseCombatWeapon	
WEA	CloseCombatWeapon	BladedWeapon
WEA	CloseCombatWeapon	PoleWeapon
WEA	Cyberweapon	
WEA	ExplosiveDevice	
WEA	ExplosiveDevice	Bomb
WEA	Gun	
WEA	Gun	Pistol
WEA	Gun	Rifle
WEA	Gun	Shotgun
WEA	Missile	
WEA	Missile	AirToAirMissile
WEA	Missile	AirToGroundMissile
WEA	Missile	AntiSatelliteMissile
WEA	Missile	ICBM

WEA	Missile	SurfaceToAirMissile
WEA	Missile	SurfaceToSurfaceMissile
WEA	NonLethalWeapon	
WEA	NonLethalWeapon	Baton
WEA	NonLethalWeapon	BatonRound
WEA	NonLethalWeapon	DirectedEnergyWeapon
WEA	NonLethalWeapon	ElectroshockWeapon
WEA	NonLethalWeapon	GasSpray
WEA	Projectile	
WEA	Projectile	Bullet
WEA	Projectile	Shell
WEA	WeaponsSystem	
WEA	WeaponsSystem	MissileSystem
WEA	WMD	
WEA	WMD	Bioweapon
WEA	WMD	ChemicalWeapon
WEA	WMD	NuclearWeapon