RECOGNYZE ANNOTATION GUIDE

Abbreviations

URI = Uniform Resource Identifier (basically a URL)

KB = Knowledge Base

Currently we are only focused on several major types of entities:

- Organizations
- People
- Location
- Events
- Products
- Works
- Misc

CLASS: ORGANIZATION

Usual subclasses: company, airline, educational institute, fraternity, sports league, sports team/club, terrorist organization, government agency, government, political party, military, police, intelligence agency, news agency, publishing house, etc.

Polisemy

Abbreviations of all kinds especially are hard to disambiguate.

Example: Bank – (1) financial institution; (2) the building where the financial institution resides; (3) to rely upon someone. Only (1) will be marked as an Organization.

Spelling Variations

It is recommended to annotate all spelling variations.

We should do like GATE: disambiguate long form.

Long and short names

See: Stanford and Stanford University.

Both should be marked with the same URI: http://www.dbpedia.org/page/Stanford_University.

Missing terms or particles in the name of the organization

Arab Banking is an old bank. Abu Sayyaf plans another bombing. (Arab Banking Group & Abu Sayyaf Group) Do consider missing terms or particles!

Branches of the same company

Sony Music Entertainment and Sony - here we should mark Sony Music Entertainment if the article refers to it.

Sony Switzerland and Sony - again we should mark Sony Switzerland if it exists, otherwise select the parent!

Product and company have (almost) the same name (company name might be incorporated)

See: Facebook, Google.

If product is synonymous with the company (Facebook) put it as the name of the company. If the name of the product is not the same than don't consider it a mention of the company.

Acronyms or Abbreviations (Currencies are removed)

See: NBA, NFL, but also abbreviations for Cantonal Banks, Universities, Sport Teams, etc.

They should be considered as an entity, as you would do it in Technical Sciences.

Multiple abbreviations for a single source

See: MLB Advanced Media (http://en.wikipedia.org/wiki/Major_League_Baseball_Advanced_Media) frequently appears with various spellings like MLB, MLB Advanced Media, Major League Baseball Advanced Media, MLBAM, mlbam

In such cases it is ok to ground to the company name (please use the long form so that all instances are grounded to the same company/organization).

Ticker symbols

GOOG, MSFT, FB should be grounded to company names.

Nested entities (TOTAL OR PARTIAL CONTAINMENT)

See: University of Vienna.

Ground just the entity that corresponds to the type of the query.

Related businesses that share the same name

There is a possibility that an organization shares a name with a parent organization without being a branch. A good example is in motor sports.

Ferrari won a new Formula 1 constructor title (Organization, Scuderia Ferrari).

Ferrari expands in Africa. (Organization, Company)

Company and product

In general try to distinguish between a company and its product, even if the product bears the same name.

He uses Facebook all day. (Product)

Facebook bought Whatsapp for a huge amount of money. (Organization)

General versus local

Some types of organizations are called by the same name regardless of the state or city they are based in. Some examples: Police, Senate, Parliament.

In these cases one has to carefully consider if the organization has a local branch and should annotate that local branch whenever possible.

The named branch can easily qualify as a named entity (Senate of Berlin, Senat von Berlin), whereas the general organization might not (He was chased by a bunch of Police officers. They met in front of the Senate. The Senate Inquiry was unnecessary long.).

NIL Candidates

When the candidate is not in the KB we consider it NIL (we do not add any URI to them).

If they have some time, annotators are asked to provide some details about the NIL entity.

CLASS: PEOPLE

Usual subclasses: actor, architect, artist, athlete, author, coach/trainer, director, doctor, engineer, monarch, musician, politician, professor, religious leader, secret agent, soldier, terrorist

Polisemy - Different people, same name

Search for *Winston Churchill* in Wikipedia, for example.

We mark each person with the corresponding URL.

Titles

In general political and university titles (those awarded to you when you finished a university cycle) should be marked as part of the name: *King George, Prof. Ben Shneiderman*.

This is especially important for German language.

Please do not mark as part of a name informal titles like: *World Champion, Olympic Medalist, Grand Slam Winner, Academy Award Winner.*

Please provide a comment where you consider necessary to mark the title as part of the name and this case is not included in the accepted cases (political titles and university titles).

People that appear mentioned just by first or last name (PART-OF)

President Bush attended the opening ceremony of the Olympic Games in Beijing. (Ziqizhang PhD thesis)

Agnetha, Björn, Benny, and Anni-Frid formed Sweden's most successful pop music group. (example from AIDA - the 4 people are known to have been members of ABBA).

If it is possible to determine who that person is, ground to the right entity in the Gold Standard, otherwise just mark it as a person, but don't ground it to anything.

Indirect references

Sweden's most successful group most probably points to http://dpedia.org/resource/Abba.

Mark these only if you consider that they are common knowledge for most humans.

Names with shortened forms

Osama Bin Laden vs. Bin Laden.

Alex vs Alexander (they might refer to the same person in the same text so it is ok to ground it to the full entity – however if we talk about Berlin Alex can also signify Alexanderplatz, a geographical location)

It is ok to ground to the long name.

Names with alternate spellings

See: Osama vs. Ussamah vs. Oussama.

Ground to the correct entity. Same like with diminutives.

Aliases

See: Osama Bin Laden vs. Sheikh Al-Mujahid

Same like President Bush rule.

Diminutives

See this page: http://en.wikipedia.org/wiki/Diminutive

Becks and Posh went to a wedding. (Translation: David Beckham and his wife Victoria went to a wedding.) It is quite clear that these will not be in the datasets, but a human user might recognize them. Ground to the correct entity if you are able to recognyze it.

Nicknames

While nicknames are not associated only with people (animals can have nicknames too), when it comes to people, it is actually common to use the nicknames as direct or indirect references.

There are nicknames that need no introduction for humans: King of Rock'n'Roll (Elvis Presley), Queen of Soul (Aretha Franklin), Queen B (Beyonce), The Fresh Prince (Will Smith), and so on.

See the following link for more details: http://en.wikipedia.org/wiki/Lists_of_nicknames

Nested entities

Bobby Sevilla - Disambiguate according to case, but Sevilla shouldn't be a location.

It is common practice in many countries to see city names as family names.

Former names

They are valid. Consider this: the former wife of a celebrity changes her name after divorce. Take the current URL of the respective person. Most likely the old name should be a redirect to the current name.

NIL Candidate.

When the candidate is not in the KB we consider it NIL.

If there is time the annotators are asked to provide as much details about the NIL entity as they can find (e.g., if Barack Obama would not be in Wikipedia/DBpedia additional details could provide data about his current job, political orientation, etc).

CLASS: LOCATION

Usual subclasses: city, country, county, province, railway, metro, road, bridge, body of water, island, mountain, glacier, astral body, cemetery, museum, park, building, airport, dam, hospital, hotel, library, power station, restaurant, sports facility, theater, subway, historical locations

Polisemy

It is quite usual to find the same name for institutions (Organizations) and their buildings (Locations). Please try to differentiate according to the situation.

Examples:

They met in front of the National Library of Medicine. (Location – Building)

The National Library of Medicine acquires all the important medical journals. (Organization)

Same name for multiple geographical entities

Same name can be applied to an entity from different locations (villages from different counties).

City names can often correspond to region names.

Please mark according to the case.

NIL Candidates

When the candidate is not in the KB we consider it NIL.

Historical Geographical Entities

They should be marked as location as well, even though they don't exist anymore. Even if geographical KBs (Geonames or others) will not have an entry on them, the text will probably make a reference to it because it either a) belongs to a recent era; b) it is a historic text; c) it can be a review of a movie/book/work that happened in that geographical space some time ago.

Would be good to try to specify the period for historical geographical entities in the comments.

CLASS: EVENT

Usual subclasses: attack, election, protest, military conflict, scandal, sports event, terrorist attack, war, classic holidays, concerts, launches

Anything that happens in the world can actually be marked as an event, but it is recommended to mark events that might be identified if you search for them online. Examples:

- Queen's New Year's Eve London Concert
- Eurovision Song Contest
- UEFA Champions League Final
- Syrian Civil War
- Second World War
- Battle of Damascus (2012)

It is also acceptable to mark as events the following:

- Anniversaries – 10th Anniversary of RBB, 50th Anniversary of Kennedy's Visit in Berlin
- Gala Dinners
- Local Events – Groundhog Day, 1st of May, Halloween
- Elections – The UK Elections from May 2015

Generic Events

Generic events (Christmas, Easter, etc.) can be marked without any mention of the year if such a mention is not available in the text (e.g., Christmas 2015).

Different Names for the Same Event

All the surface forms for an event should be marked with the corresponding entity.

Battle of Damascus and *Operation Damascus Volcano* could therefore represent the same event if the text refers to the *Battle of Damascus (2012)*. Note that multiple battles were recorded in the area in different years.

Different Years Policy

If the Event represents a generic holiday (1st of May, Christmas) just mark it as such. If the Event represents a Political Event, War, Riot or something similar that might have happened repeatedly, mark the events that correspond to the year in the text, or that might be closest in relevance to the events described in the text.

Same Year, Different Dates Policy

Sometimes multiple events can take place in the same year in the same location and be labelled similarly. *The Paris Attack 2015* can refer to the *Charlie Hebdo shooting* and to the *Bataclan shooting*. Please disambiguate accordingly (try to mark the event described in the text).

Event Identification Policy

Please only mark events that can easily be identified (e.g., London Riots, September 11, etc.).

Do not mark things that sound like events (e.g., a mention of a concert) if there is no additional information to place them in a certain time and location (e.g., the Eagles of Death Metal concert from Bataclan, Paris, November 13). There is no need for the information about the event to be complete, but it should allow for at least a fuzzy identification of the respective event. Keep in mind that a mention to a random event (AC/DC concert) can refer to many similar events if there is no additional information.

CLASS: PRODUCT

Usual subclasses: engine, airplane, car, ship, spacecraft, train, camera, phone, computer, software, game, instrument, weapon, magazine, newspaper, social network, food, brand

Anything that is created for mass production should be considered **Product**, except the cases in which it might be a work of art (television series, book, album, painting, and so on).

Some example of products:

- Facebook (social network)
- Google (social network)
- Porsche 911 (car)
- Uncle Ben’s White Rice (food)
- Big Mac (food)
- Apple (brand)

Product and Company Name

If the name of the company and its product name are the same (e.g., *Facebook*, *Google*), annotate according to the given context (e.g., *Facebook* social network if the text mentions social network).

Generic Products vs Brands

Try to annotate brands instead of generic products. Generic products should not be marked at all.

Example: *Mars chocolate* is better than just annotating *chocolate*.

Specificity – Exact Reference

If the product is specific, please mark it as such, instead of marking the brand.

Example: *Porsche 911* should be marked instead of *Porsche* if it is present in the text.

CLASS: WORK (OF ART)

Usual subclasses: film, play, TV show (including TV series), written work, music, entertainment, sculpture, painting, book, game

In general classic music, classic books, sculptures can be considered **works of art**.

In general works of art have a recognized author (and it is recommended that you mark them together with their author if the author appears):

- Michelangelo's David (sculpture)
- Da Vinci's Mona Lisa (painting)
- Gutenberg Bible (book, but Gutenberg is actually the printer and editor rather than the author)
- Shakespeare's sonnets (book)
- Bram Stoker's Dracula (book)
- Ravel's Bolero (music)
- Need for Speed
- FIFA 2015 (the video game)

Mention of Authorship

In general mentions of classic works can be accompanied by the name of their creator (e.g., Da Vinci's Mona Lisa, Ravel's Bolero). If this happens please include author's name in the annotation.

The author of a work of art can also be anonymous, especially if that object is really old (e.g., The Bible, books with unknown authors).

Episode, Chapter, Fragment, Aria

If a fragment of a known work is identified (episode, chapter, aria, video fragments, etc.) please mark it accordingly. It is ok in such situations to mark both the fragment (episode) and the parent work (television series) if both of them are available in the text.

Works that are not mentioned directly by their name

The Scripture obviously would point to The Bible so mark it as such. If any such case is met proceed in a similar manner.

CLASS: MISC

Literally anything else that can be classified as a Named Entity and doesn't fit in the previous classes.

We are mostly interested in the previous classes in this evaluation (especially since we are focused on media monitoring), but it is important to know what else should be marked.

CROSS-TYPE DISAMBIGUATIONS POLICY

If the entity is highly ambiguous (*Sherlock* can be a fictional character, TV show, book, person name, etc) please disambiguate according to the context (perhaps if Benedict Cumberbatch the text clearly refers to the BBC TV show).