Annotation Guidelines

This text provides annotation guidelines for the anonymisation of Wikipedia summaries with the Tagtog tool.

You are given a collection of Wikipedia summaries (in a directory with your name). The goal of your annotation is to mark all text spans (= a continuous stretch of one or more words) that correspond to information tied to the individual and, afterwards, indicate whether they can be used to re-identify the individual to be protected or not.

The data has been *pre-annotated* automatically. Your annotation work consists of four stages:

- **Step 0:** Read through the summary once
- **Step 1: Entities.** Taking the pre-annotations as a starting point, annotate each entity with a semantic type (PERSON, CODE, LOC, ORG, DEM, DATETIME, QUANTITY, MISC, see below), remove/correct errors in the pre-annotated spans, and manually annotate spans that went undetected.
- **Step 2: Masking.** For each span annotated in Step 1, specify their identifier type (direct identifier, quasi-identifier, or does not need masking). See below for detailed definitions of these terms.
- **Step 3: Final review.** Save and confirm your annotation, and inspect the review document generated in your `masked` subfolder to ensure you have not missed any attributes Otherwise, go back to Step 1 or 2.

Step 1: Pre-annotated spans

Step 1 focuses on correcting/removing pre-annotations that were generated automatically. *Those pre-annotations are just a starting point: you must actively seek to correct and extend the annotation*. This means that you should always:

1. Verify whether each pre-annotated span is correct or needs to be removed or edited (e.g. whether the span should be changed).
2. Manually annotate each span that includes information tied to the individual but was not detected by the automatic process.
3. If two mentions refer to the same underlying entity but have a different string (e.g. "Bernie Brennan" and "Brennan"), insert a reference *relation* between them (see below).

For Step 1, you do not need to worry about *re-identification risks* (this will come in Step 2), you just need to mark all span.

The texts that will be used for the annotation are the first paragraph of Wikipedia articles, where one can find a small but informative summary of the individual. Wikipedia summaries, though usually shorter in size are very rich with information (mainly quasi identifiers), so there is not a lot of text that should be left non-annotated at this Step. The most usual error that can be found in the pre-annotations provided are boundary issues (e.g. George W. Bush to be corrected to George W. Bush) .

Entities that occur several times through a document are visually marked in TagTog with a yellow border for all mentions except the first one (to indicate that the subsequent mentions are "derived" from the first one). When an entity needs to be corrected or removed, you should

therefore apply your changes on the first mention - this way, your changes will be automatically propagated to all mentions, without having to repeat your corrections one by one.

The list of categories is presented below. If the information does not fit within one of these, the MISC category should be used.

Category	Description
PERSON	Names of people, including nicknames/aliases, usernames and initials
LOC	Places and locations, such as: • Cities, areas, countries, etc. • Addresses • Named infrastructures (bus stops, bridges, etc.)
ORG	Names of organisations, such as: • public and private companies • schools, universities, public institutions, prisons, healthcare institutions non-governmental organisations, churches, etc.
DEM	Demographic attribute of a person, such as: • Native language, descent, heritage, ethnicity • Job titles, ranks, education • Physical descriptions, diagnosis, birthmarks, ages
DATETIME	Description of a specific date (e.g. October 3, 2018 or 2018-10-03), time (e.g. 9:48 AM) or duration (e.g. 18 years).
QUANTITY	Description of a meaninful quantity, e.g. percentages, monetary values, meters, miles etc.
MISC	Every other type of information that describes an individual and that does not belong to the categories above

Examples

In general, annotation should mark the *minimal span* which denotes the entity or property in question.

PERSON

For person names the annotation should include suffixes, prefixes and titles, such as Sr., Jr., Sir etc.

- Robert John Downey Jr.
- Sir Elton Hercules John
- Robert Conner Sr.
- Princess Bona of Savoy-Genoa
- Lieutenant General Faridoon Noshir Billimoria
- Clive Adrian Stafford Smith OBE

Some examples which are regarded as names:

- Names: E.g., ‘Harry Hole’, 'Hole', 'Harry'
- Initials: E.g. 'H.H.'

- Spelling mistakes: E.g. 'Hary Hole'
- All orthographic variations E.g. 'harry hole', 'Harry HOLE'
- Nicknames, aliases, usernames E.g. 'haryh, 'Stasis'

Translations of the name should be separate spans, but remember to add the relation between the spans:

- Ikuo Takahara (Japanese: 高原 郁夫)

NB: The name of a person can be both a direct identifier (if it is the full name of the individual to protect) or a quasi-identifier (if it is the name of another related person).

DATETIME

All parts of a date should be included in the same span. Separate entities connected by "-", "and", "or", etc, should be annotated separately if they refer to different events in time (e.g. date of birth and date of death). Prepositions (e.g. on, at) should not be included in the same span:

- March 23, 1987
- 1987-2001
- born in 1987

Exception If the dates partially overlap because they refer to a continuous event in time (e.g. a band tour) they should be part of the same span:

- 10 and 12 of March 1987

Definite or indefinite articles are typically not included in the span, unless explicitly part of the entity (as for titles of books and movies, such as “The Great Gatsby”).

- member of the Republican Party of Minessota
- according to the United States House of Representatives
- she wrote The Nightingale

LOC

Includes cities, areas, counties, addresses, as well as other geographical places, buildings and facilities. Other examples are airports, churches, restaurants, hotels, tourist attractions, hospitals, shops, street addresses, roads, oceans, fjords, mountains, parks.

- Austria
- Orange County
- Mount Kimbie

Separate entities connected with a conjunction (e.g. ‘and’) should be annotated separately:

- he visitied Los Angeles and New York (LOC)

In case an entity could be either ORG or LOC (e.g. Turkey or Breitvet prison), the entity type best describing the referent should be chosen, i.e. ORG if the occurrence refers to the institution itself, and LOC if it refers to a geographic location.

DEM

These are demographic properties and include both physical, cultural and occupational/educational properties, such as various physical descriptions, diagnosis, native language, ethnicity, job titles, age, etc.

- 40 years old
- he was a journalist
- a group of left-wing extremists
- diagnosed with motor neurone disease
- a Polish and naturalized-French physicist

Pronouns (he, she) should not be annotated to protect gender information.

QUANTITY

Units, such as currencies, should be included in the **same span**.

- $37.5 million
- 375 euros
- 4267 SEK
- 1000 Kilos
- 4 meters

But consider the following example :

"Lorenzo Smith is an American singer-songwriter who has released three albums"

that could be annotated as :

"Lorenzo Smith is an American singer-songwriter who has released three albums"

where by masking the quantity of the albums we are left with a good trade-off between privacy and data utility. In cases like these the quantity should not be part of the span.

MISC

Other (quasi-)identifying words such as trademarks, products, events, etc. All things artificially produced are regarded products. This may include more abstract entities, such as speeches, radio shows, programming languages, contracts, laws and even ideas (if they are named).

- founding member
- punk-rock

Brands are products when they refer to a product or a line of products, but organisation when they refer to the acting or producing entity.

Relations

Finally, if some text spans are referring to the same underlying entity through different mentions (such as “Forrest Carlisle Pogue Jr.” and “Pogue”, or “Steve Pickton” and “Stasis”), annotate those referential relations. More precisely:

- a. Find an occurrence of the first mention (such as “Forrest Carlisle Pogue Jr.”) and click “add relation”
- b. Find an occurrence of the second mention (such as “Pogue”) and click on it. You should now see a relation between the two.

You do not need to add “same_as” relations between entities with the exact same string. It is also sufficient to annotate the relation once, even though there may be several occurrences of both mentions.

Step 2: Masking

In this stage of the annotation, you should review all text spans marked in Step 1, and specify whether they need to be masked (either as a DIRECT or QUASI identifier) to protect the identity of the individual specified in the annotation task. You should mark all direct and quasi identifiers but *not more than those* (we still wish to retain as much textual content as possible).

For every entity annotated during stage 1, you should set the correct value for the following: **identifier_type**

- DIRECT_ID (direct identifiers): text spans that *directly* and *unequivocally* identify the individual to protect in the case and should therefore be masked.
- QUASI_ID (quasi-identifiers): text spans that should be masked since they may lead to the re-identification of the individual when combined with other (not masked) quasi-identifiers mentioned in the text along with public background knowledge.
- NO_MASK: entities that are neither of the above, and should therefore not need to be masked. Start by applying NO_MASK to the least precise/specific quasi-identifiers in the summary first.

Since as mentioned the texts are usually shorter in size, make sure to, in Step 3, mask as little as possible while keeping the individual protected. We wish to retain as much of the text as possible, without leaving 'dangerous' entities in the text.

We recommend you use TagTog's Document review mode (press r) to easily go through all entities one by one without having to click on anything. When deciding on the identifier type and confidential status of an entity, make sure you select the first mention - this way, your decision will be automatically propagated to all subsequent mentions of that same entity (shown with a yellow border on TagTog).

Direct identifiers

= text spans that contain information that directly and unequivocally identify the individual to be protected.

Examples: person names (including nicknames/aliases and usernames), no matter where they appear in the text.

Quasi-identifiers

= Information that, in isolation, does not identify the individual to be protected but can do so, in combination with other quasi-identifiers and background knowledge. These will often refer to demographical (“an Austrian poet”) or spatiotemporal attributes (“on February 6 in Sevilla”). For instance, the combination of date of birth, gender and profession will typically allow you to find out the identity of a person.

For a re-identification to be possible, quasi-identifiers must refer to some information that can be seen as potential “publicly available knowledge” — i.e. something that we can expect that an external person may already know on the individual or may be able to find —, and the

combination of quasi-identifying information should be enough to re-identify the individual with no or low ambiguity. You should judge whether it is likely that someone could, based on public knowledge, know the quasi-identifying values of the individual to be protected. There is some room for interpretation here, but the annotator should ask themselves the question: if I wanted to find out the identify of the individual in the document, should I expect to be able to connect those pieces of information with some other knowledge sources (such as news articles, social media, census data, etc.)? and, are those pieces of information enough to re-identify the individual with no or low ambiguity? *In the vast majority of cases, you don't actually need to do any search for those knowledge sources, your intuition will suffice.*

As a rule of thumb, immutable personal attributes (e.g., date-of-birth) on an individual that can be known by external entities should be considered quasi-identifiers. Circumstantial attributes may be considered quasi-identifiers or not according to the chance that external entities may know such information (e.g., current place-of-living or a hospital admission date could be, but the number of times one has gone to the grocery store in a week may not). If you feel like looking up a piece for information to see if it would be considered background knowledge, before deciding on whether to MASK a span or not, make sure to ignore any site Wiki-related (e.g. Wikipedia, Wikidata, Wikimedia etc.) since these are the source of the data and the information in the pre-annotations.

Usually, only very general attribute values that encompass a large number of individuals (e.g., country-of-birth) may be ignored, since they would match a large population of individuals and would not enable a unequivocal re-identification. This also depends on the presence of other quasi-identifiers within the same document: the larger the amount and the more concrete the information they provide, the larger the chance that they may enable re-identifications.

Step 3: Final Review

Save and confirm your annotation. After saving the document, a new "review" document should be now be available in your subfolder `masked`. For instance, if your document lies in the path `yourname/text1`, a new document should now appear at `yourname/masked/text1`. (*if that's not the case, let us know*)

This document shows which entity will end up being masked in an anonymised version of the summary. Direct identifiers and quasi-identifiers are now replaced by *****. Confidential attributes are shown in clear text but marked as entity. Inspect the text to ensure you haven't forgotten any (direct or quasi) identifier or confidential attributes. Otherwise, go back to step 1 and 2 and repeat the process.

A step by step example 1:

- **Step 0:** Read through the summary once

Pavel Bobek(PERSON) (16 September 1937(DATETIME) – 20 November 2013(DATETIME)) was a Czech singer(DEM).

- **Step 1: Entities.** Taking the pre-annotations as a starting point, remove/correct errors in the pre-annotated entities, and manually annotate entities that went undetected.

Pavel Bobek(PERSON) (16 September 1937(DATETIME) – 20 November 2013(DATETIME)) was a Czech(DEM) singer(DEM).

- **Step 2: Masking.** For each entity annotated in Step 1, specify whether their identifier type (direct identifier, quasi-identifier, or does not need masking) and their confidential status.

Pavel BobekDIRECT (16 September 1937QUASI – 20 November 2013QUASI) was a CzechNO–MASK singerNO–MASK.

- **Step 3: Final review.** Save and confirm your annotation, and inspect the review document generated in your `masked` subfolder to ensure you have not missed any identifier or confidential attribute. Otherwise, go back to Step 1 or 2.

A step by step example 2:

- **Step 0:** Read through the summary once

Lorenzo Smith(PERSON) (born May 23, 1972(DATETIME)) is an American(DEM) singer-songwriter(DEM) who has released three(QUANTITY) albums(MISC).

- **Step 1: Entities.** Taking the pre-annotations as a starting point, remove/correct errors in the pre-annotated entities, and manually annotate entities that went undetected.

Lorenzo Smith(PERSON) (born May 23, 1972(DATETIME)) is an American(DEM) singer-songwriter(DEM) who has released three(QUANTITY) albums(MISC).

- **Step 2: Masking.** For each entity annotated in Step 1, specify whether their identifier type (direct identifier, quasi-identifier, or does not need masking) and their confidential status.

Lorenzo SmithDIRECT (born May 23, 1972QUASI) is an AmericanNO–MASK singer-songwriterNO–MASK who has released threeNO–MASK albumsNO–MASK .

OR

Lorenzo SmithDIRECT (born May 23, 1972QUASI) is an AmericanQUASI singer-songwriterNO–MASK who has released threeNO–MASK albumsNO–MASK .

OR

Lorenzo SmithDIRECT (born May 23, 1972QUASI) is an AmericanNO–MASK singer-songwriterNO–MASK who has released threeQUASI albumsQUASI .

- **Step 3: Final review.** Save and confirm your annotation, and inspect the review document generated in your `masked` subfolder to ensure you have not missed any identifier or confidential attribute. Otherwise, go back to Step 1 or 2.