Impresso Named Entity Annotation Guidelines

Version: 2.2. - January 2020

Maud Ehrmann, Camille Watter, Matteo Romanello, Simon Clematide[1]

These guidelines were primarily written to support the manual annotation of *impresso* evaluation corpus. They are however also useful for system developers participating in the HIPE shared task. In this case, instructions regarding manual annotation concrete instructions can be ignored.

[1] Camille Watter for initial Quaero translation and impresso adjustments, Maud Ehrmann for reshaping, reformulation and impresso adjustments, Simon Clematide and Matteo Romanello for impresso adjustments.

1. Preamble

Guidelines genealogy

Impresso NE annotation guidelines are derived from Quaero guidelines[2]. Originally designed for the annotation of "extended" named entities (i.e. more than the 3 or 4 traditional classes) in French speech transcriptions, Quaero guidelines have furthermore been used on historic press corpora[3]. *Impresso* guidelines main's difference with respect to Quaero's is *reduction*: only a subset of Quaero entity types and components are considered, as well as a subset of linguistic units eligible as named entities. These adaptations result from what we deemed most relevant to annotate in our context, and from time and resource constraints. Despite these adaptations, *impresso* annotated corpora will mostly remain compatible with Quaero guidelines.

Application context

The objective is to extract information from historical newspaper articles, in view of supporting the search, filtering and analysis of large collection of newspaper archives, and of building a historical knowledge base, eventually connected to others (e.g. Wikidata, HistHub).

As such, our objective is similar to one of classical media monitoring, where we want to extract salient 'journalistic' entities among the typical '5Ws' (Who, What, Where, When, Why).

Our context is however different in that documents are not contemporary but historical, and final users are not politicians or economic actors but scholars. This led us to some adjustments with respect to, mainly: (a) the tag set (addition of newspaper-related specific types), (b) granularity of annotation (emphasis on *Person* type in view of the biographical scenario), and (c) concrete implementation of annotation (flag for noisy entities, capacity to view the original facsimile).

Methodology

These guidelines were built as follows:

1. translation of Quaero guidelines from French to English;
2. selection of high level types of interest during a workshop with historians;
3. constitution of "mini-reference" corpora from *impresso* sources for French and German;
4. annotation of mini-references to test and validate our guidelines, with several iterations of the following process: [annotation => collection of problems => discussion and arbitration => adaptation of guidelines => annotation];
5. curation of mini-references, selection of exemplary difficult cases;
6. final validation of guidelines.

We use INCEpTION as annotation tool, with the visualisation of image segments.

Status

Guidelines are considered as stable but minor changes *could* happen.

[2] See the original Quaero guidelines: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf , and our English translation: https://docs.google.com/document/d/13LRvP5Qh99myEEH_lqqcHaa3S-nZ2Sr71iZ5YbecDCc/edit#

[3] See ELRA catalog entry: http://catalog.elra.info/en-us/repository/browse/ELRA-W0073/

2. General instructions

2.1 Entity types and subtypes

The objective is to annotate all named mentions in texts, of the following types and subtypes:

Type	Subtypes	
pers	pers.ind pers.coll pers.ind.articleauthor	
org	org.adm org.ent	org.ent.pressagency
prod	prod.media prod.doctr	
time	time.date.abs	

Type	Subtypes	
loc	loc.adm	loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup
	loc.phys	loc.phys.geo loc.phys.hydro loc.phys.astro
	loc.oro	
	loc.fac	
	loc.add	loc.add.phys loc.add.elec
	loc.unk	

Additionally, types can be complemented with the following flags:

- unresolvable
- noisy entity
- literal

A. Most specific label. Manual annotation uses subtypes only (e.g. `<pers.ind>`). It is not allowed to use a type (e.g. `<pers>`).

B. Unknown. When it is not possible to figure out the subtype, one can use `unknown` (`.unk`). In *impresso* this subtype is available for locations only. Note that this is different from ambiguity, whose annotation is explained in section 1.3.

2.2 Named entity mention lexical characteristics

A. Nature.

Linguistic units considered as named entities must include a proper name, or a definite description having the status of a proper name[4]. Although the definition of a proper name is not straightforward, here are a few characteristics commonly accepted (not valid in all cases nor in all languages): presence of majuscule, non inclusion in lexical but in encyclopedic dictionaries, absence of meaning (the name *George* does not carry - per se - any information about the type of entity that can be called this name, while the noun "table" gives specific information about the type of objects that can be called by it - i.e. having a plateau and feets), and absence of compound meaning (the *White House* does not refer to any house which is white, la *Gare de Lyon* is not in Lyon, le *Pont Neuf* is very old).

We do not specify further the definition of proper names[5], but instead rely on the linguistic intuition/awareness of annotators, who should always keep in mind our objective of extracting 'journalistic' information typically conveyed via referential entities. There will be borderline cases, which we ask annotators to report in a separate file for further discussion[6].

Phrases such as

- *Die präkolumbianische Zivilisation, la civilisation précolombienne*
- *l'armée bavaroise*
- *les forces tchadiennes*
- *le gouvernement français*

are *not* annotated because they do not contain proper names.

Phrases such as:

- *le gouvernement Franco*

  le <org.adm> gouvernement
  <comp.name> <pers.ind> Franco </pers.ind> </comp.name>
  </org.adm>

are annotated.

In front of some definite descriptions, it might be difficult to decide what to do, e.g. *la commission Impériale, l'escadre de Nelson.* In such difficult cases, consider the following:

- definite descriptions which can be considered as named entities tend to have a **nominative function** (like proper names) rather than a descriptive function. What a definite description says literally about a referent is less important than the nominative aspect.
- even though, some named entities are definite descriptions which are descriptive, e.g. "Syndicat National de la Magistrature". In such cases, what makes it a named entity is the **referential stability**: the entity referred to is always the same.

[4] This position is more strict than Quaero, which allow entities to be composed of proper names and of common nouns (cf. Section 1.5 or Quaero guidelines).
[5] A rabbit hole. For an overview of proper name definition see: https://hal.archives-ouvertes.fr/tel-01639190
[6] See the last section "Quick guide and concrete implementation".

- in general, our bottom line is: **we do not accept borderline definite descriptions.**

B. Boundaries. A named entity can be the head of several nominal syntagms but not all of them are annotated.

- Named entity mentions exclude:
  - subordinate clauses;
  - incidental clauses or insertions : if an insertion divides a mention, each part is annotated separately;
  - determiners.

- Named entity mentions include:
  - pre modifiers
    Le soviétique Alexandre Avreni a déclaré…
    Le compatriote Serge Martin est déçu…
    La grande Armée Rouge

  - post modifiers, including in apposition:
    Anne Hidalgo, maire de Paris, a déclaré
    Anne Hidalgo, une forte femme, a déclaré
    Shekau, chef de l'une des trois factions de Boko Haram et fondateur historique du groupe, diffusait une vidéo…

- For time expressions, the left bound includes everything preceding the expression in the form of a nominal or a prepositional group. Articles and determiners are included for dates (unlike for other entities). Prepositions and adjectives are therefore part of date expressions.
  … *der 23. Januar 1826*…
  … *le 23 janvier 1826*…

- Special cases with noisy OCR:
  When it is difficult to establish the boundary of a mention because of noisy OCR:
  - look at the image
  - include, in the annotation, the garbage characters which you think should have been recognized and should be part of the mention
  - mark the mention with the flag "noisy-entity" and add your OCR hypothesis correction.
    ex: in the string *Trève ** (which stands for *Trèves*), the full string *Trève ** should be annotated, not only *Trève.*

- Special case with German compounds:
  Apply the cross-lingual or decomposition test, i.e. translate the compound to French and in the German compound annotate only what should be annotated in French.

Baslerpropaganda
=> French translation (decomposition): propagande bâloise
=> no annotation

Zürichputsch
=> French translation (decomposition): le putsh de Zurich (Putsch von Zürich)
=> annotation of "Zürich"
<loc.adm.town>Zürich</loc.adm.town>putsch

Donaufestungen
=> Festungen an der Donau
=> annotation of "Donau"
<loc.phys.hydro>Donau</loc.phys.hydro>festungen

Der am Montag in Kairo ermordete ägyptische Ministerpräsident Al-Nokraschi
=> "Le premier ministre égyptien Al-Nokraschi, qui a été assassiné au Caire lundi, ..."
<pers.ind>
<comp.func> <comp.demonym>ägyptische</comp.demonym>
Ministerpräsident</comp.func>
<comp.name> Al-Nokraschi</comp.name>
</pers.ind>

The connecting "s" in German compounds is *not* annotated:
Völkerbundsmitgliedern
=> only *Völkerbund* is annotated
<org.adm>Völkerbund</org.adm>smitgliedern

2.3 Components

A named entity mention consists of one or more components, as well as parts without explicit component markup. *Impresso* partially follows Quaero's guidelines and consider the following components:

A. Component of the type *Person*:

- comp.func
- comp.title
- comp.name
- comp.qualifier
- comp.demonym

B. Component for all types except *Time:*

- name, used to mark the name of the entity.

The component name is optional when the mention contains only one name:
<loc.adm.town> Paris </loc.adm.town> (name optional)
la <pers.ind> maire de Paris <comp.name>Anne Hidalgo</comp.name></pers.ind>

2.4 Nesting and special constructions

A. Nested entities. An entity can be nested in another entity or in an entity component.

- nested entities are annotated for the types PERS, LOC, ORG, without limit of nesting level during the annotation phase. For system evaluation, nested entities are only of depth 1.

La Feuille d'Avis de Neuchâtel
<prod.media>Feuille d'Avis de
<loc.admin.town>Neuchatel</loc.admin.town>
</prod.media>

La société du Parc du Creux-du-Vent...
<org.adm>société du
<loc.oro>Parc du <loc.phys.geo>Creux-du-Vent</loc.phys.geo>
</loc.oro>
</org.adm>

Le maire de Paris Bertrand Delanoë a déclaré
<pers.ind>
<comp.func>maire de <loc.adm.town>Paris </loc.adm.town>
</comp.func>
<comp.name>Bertrand Delanoë</comp.name>
</pers.ind>

dem Preussischen Staatsminister der auswärtigen AngelegenHeiten, Graf von Goltz
<pers.ind>
<comp.func> <comp.demonym>Preussischen</comp.demonym>
Staatsminister der auswärtigen Angelegenheiten</comp.func>
<comp.title>Graf</comp.title>
<comp.name>von Goltz</comp.name>
</pers.ind>

- components of nested entities are not annotated

B. Coordination. Entities coordinated based on a common descriptor or trigger word are annotated separately. Type is inferred from the type of the coordinated entity. Coordinating conjunctions are excluded from annotation.

Der Bodensee, Starnberger See und Müritz

Der <loc.phys.hydro> Bodensee </loc.phys.hydro>,
<loc.phys.hydro> <comp.name>Starnberger</comp.name> See
</loc.phys.hydro>, und
<loc.phys.hydro> Müritz </loc.phys.hydro>

vallées de la Lorraine, de l' Alsace et de la Champagne
<loc.phys.geo> vallées de la <comp.name>Lorraine</name> </loc.phys.geo>,
<loc.phys.geo> de l' <comp.name>Alsace</comp.name> </loc.phys.geo> et
<loc.phys.geo> de la <comp.name>Champagne</comp.name> </loc.phys.geo>

In any cases, a proper name must be present in the entity mention, therefore only one entity is annotated when it is not the mentions but the title/trigger words which are coordinated:

Monsieur et Madame Chirac…
Monsieur et <pers.ind><title>Madame</title> <name>Chirac </name>
</pers.ind> …

Ost und Mitteleuropa….
Ost und <loc.admin.sup>Mitteleuropa</loc.admin.sup>....

Special case of a coordination within a component: this produces 2 separate components, excluding the coordination.

Shekau, chef de l'une des trois factions de Boko Haram et fondateur historique du groupe, diffusait une vidéo…
<pers.ind> <comp.name>Shekau</comp.name>,
<comp.func>chef de l'une des trois factions de
<org.ent>Boko Haram<org.ent></comp.func> et
<comp.func>fondateur historique du groupe</comp.func>
</pers.ind>

C. Elaboration. When a mention is complemented with an acronym or an abbreviation, both are treated as distinct entities.

DAISY das dynamische Auskunfts- und Informationssystem
<org.ent> DAISY </org.ent> das
<org.ent> Dynamische Auskunfts- und Informationssystem </org.ent>

Agipi association d'assurés pour la prévoyance, la dépendance et l'épargne-retraite
<org.ent> Agipi</org.ent>\
<org.ent> Association d'assurés pour la prévoyance , la dépendance et l'épargne-retraite
</org.ent>

D. Difficult example(s)

der bekannte Irländer Theobald Wolfe Tone, den man auf….

<pers.ind>
<comp.qualifier> bekannte </comp.qualifier>
<comp.demonym> Irländer </comp.demonym>
<comp.name>Theobald Wolfe Tone </comp.name>
</pers.ind>

2.5 Ambiguities

A. Unsolvable entity type ambiguities: flag 'unsolvable'

Even in context, some entities can remain ambiguous:

<??>Yves Rocher</??> lässt sich in Vannes nieder
<??>Yves Rocher</??> va s'installer à Vannes

In these cases, the annotation is 'double' and includes 2 types. To differentiate this annotation from a metonymic one (which also results in two tags for one mention), annotator should add the flag 'unsolvable' to one of the 2 annotations.

In case of unsolvable ambiguity, it is mandatory to indicate 2 types minimum.

B. Metonymy.

Metonymy is a figure of speech in which a thing or a concept is not called by its own name but by the name of something intimately associated to that thing or concept. The category to which the mentioned entity inherently belongs is annotated and is nested within the category that the term refers to in the context.

In Inception annotation tool, the literal annotation has to be flagged with the corresponding flag.

<org.adm>
<loc.adm.nat>*Nigeria* </org.adm.nat>
</org.adm>
weist die Anschuldigungen zurück.

Le
<org.adm>
<loc.adm.nat>*Nigeria* </loc.adm.nat>
</org.adm>
n'en a pas encore fini avec Boko Haram.

<org.adm>
<loc.adm.nat>*Frankreich* </org.adm.nat>
</org.adm>
hat den Alliierten geholfen.

La <org.adm>
<loc.adm.nat>*France* </loc.adm.nat>
</org.adm>
est intervenue auprès de ses alliés

Eine Erklärung des Quai d'Orsay
Eine Erklärung des
<org.adm>
<loc.oro> Quai d'Orsay </loc.oro>
</org.adm>

Une déclaration du Quai d'Orsay
Une déclaration du
<org.adm>
<loc.oro> Quai d'Orsay </loc.oro>
</org.adm>

Die rue de Grenelle hat auf diese Aussage reagiert
Die
<org.adm>
<loc.oro> rue de Grenelle </loc.oro>
</org.adm> hat auf diese Aussage reagiert

Die Élysée erklärt
Die <org.adm><loc.fac> Élysée
</loc.fac></org.adm> erklärt

La rue de Grenelle a réagi à cette déclaration
La
<org.adm>
<loc.oro> rue de Grenelle</loc.oro>
</org.adm> *a réagi* à cette déclaration

L'Élysée a déclaré...
L' <org.adm><loc.fac>Élysée
</loc.fac></org.adm> a déclaré...

In case of hesitation whether there should be a metonymy annotation or not, always provide the literal but be generous and consider the metonymic annotation.

For country names used on the political sense - the majority of the metonymic entities encountered in our news corpus -a possible test is to qualify the entity and see what is more natural and/or obvious: *the geographical entity X has/did*... vs. *the political entity X has/did*...

3. Entities

3.1 Person

A. Subtypes

- `pers.ind`: when the entity refers to an individual.
- `pers.ind.articleauthor`: special *impresso* type to recognize authors of newspaper articles, either full names or initials at the end of the text, or within a formula such as "*from or correspondant xx in yy*"
- `pers.coll`: when the entity refers to more than one individual.
  Even in the case of a collective person annotation, there must be the presence of a proper name (e.g. *the Beatles, the Cohen Brothers, die Habsburger, les Bourbons*).

  The following expressions are **not** annotated:

  > *die französischen Opfer des Unfalls, die chinesischen Touristen / les victimes françaises de l'accident, les voyageurs chinois*
  >
  > *Die Maya Zivilisation / la civilisation Maya*
  >
  > *Arbeiter, Menschen, die Verletzten*; / *le monde ouvrier, les êtres humains, les blessés, etc.*
  >
  > *Die Protestanten, die Spanier / les protestants, les espagnols*

B. Coverage of the type Person

- Considered as Person:
  - real persons
  - imaginary characters and characters of literature pieces (e. g. *Asterix,* when referring to the character, but not when referring to the work e.g. *Uderzo ist der Schöpfer der Comic-Reihe Asterix, Uderzo est le créateur de la BD Astérix*)
  - religious figures (*God*)

- Not considered as Person:
  - expressions which do not contain a proper name
  - demonyms which do not modify a proper name:
    e.g. Le français s'est classé quatrième.
    Der Schweizer ist Vierter geworden
  - isolated functions not attached to a person name

Der Bürgermeister von Paris => only 'Paris'	*le maire de Paris* => only 'Paris'
Die Bürgermeister von Frankreich => only 'France'	*les maires de France* => only 'France'
Der Forscher des CNRS => only 'CNRS'	*le chercheur CNRS* => only 'CNRS'
Der Präfekt ist essen gegangen => no annotation	*le préfet est parti manger* => no annotation
Der Minister für ausländische Angelegenheiten => no annotation	*le ministre des affaires étrangères*=> no annotation
Der amerikanische Minister für ausländische Angelegenheiten => no annotation	*le ministre américain des affaires étrangères*=> no annotation
Ein britischer Journalist => no annotation	*un journaliste britannique*=> no annotation
Der ehemalige Bürgermeister von Paris => only 'Paris'	*l'ancien maire de Paris* => only 'Paris'
Die Polizisten => no annotation	*les pompiers*=> no annotation
Die Polizisten von Paris => only 'Paris'	*les pompiers de Paris* => only 'Paris'
Präsident der Republik => no annotation	*président de la république*=> no annotation
Präsident der islamische Republik Pakistan => only `Pakistan'	*président de la République islamique du Pakistan* => annotate only `Pakistan'
Einer der Polizisten => no annotation	*l'un des pompiers*=> no annotation
Ex Miss Italien => no annotation	*ex Miss Italie* => no annotation
Der Papst => no annotation	*le Pape* => no annotation

C. Person Components

- **`comp.func:`**
  - an occupation (e.g. *footballer*, *negotiator*), including specialities (e.g. *Spezialist für die Jagd, Experten für Sprengstoff; spécialiste de la chasse, experts en explosifs*).

- an administrative function in public or private area (*CEO*, *minister, secretary of state*).
- 'social' roles or status, e.g. *Obdachlose, Arbeitslose, Häftlinge; SDF, chômeur, détenu.*

A function always includes the organization, place or specialization attached to it.

- **`comp.title:`** a civil or honorific address (*Mr., Mrs., Her Altess*), military titles (*amiral, général*), nobility titles (*Baron, Duc*), as well as royal titles (*King, Queen, Prince*).
- **`comp.qualifier:`** any adjective qualifying the entity: *socialiste, great, venerable*
- **`comp.name:`** covers first|middle|last|nickname
- **`demonym:`** a noun or adjective that identifies residents of a particular place

func / title / name	
Seine Königliche Hoheit Prinz Rainier <pers.ind> <title>Seine Königliche Hoheit</title> <title> Prinz </title> <name> Rainier </name> </pers.ind>	*Son Altesse Royale le prince Rainier* <pers.ind> <title>Son Altesse Royale</title> le <title>prince</title> <name>Rainier </name></pers.ind>
Der König Mohamed VI Der <pers.ind> <title> König </title> <name> Mohamed </name> <qualifier> VI </qualifier> </pers.ind>	*Le roi Mohamed VI* Le <pers.ind> <title> roi </title> <name> Mohamed </name> <qualifier> VI </qualifier> </pers.ind>
Ihr Majestät der König Mohamed VI <pers.ind> <title>Ihre Majestät </title> der <title> König </title> <name> Mohamed </name> <qualifier> VI </qualifier> </pers.ind>	*Sa Majesté le roi Mohamed VI* <pers.ind> <title> Sa Majesté</title> le <title> roi </title> <name> Mohamed </name> <qualifier> VI </qualifier> </pers.ind>
Der Dr. Duboc, ehemaliger Abteilungsleiter von Pitié-Salpêtrière Der <pers.ind> <title> Dr. </title>\ <name> Duboc </name> <func> ehemaliger Abteilungsleiter von Pitié-Salpêtrière </func> </pers.ind>	*Le Dr. Duboc, ancien chef de service à la Pitié-Salpêtrière* Le <pers.ind> <title> Dr. </title>\ <name> Duboc </name> <func> ancien chef de service à la Pitié-Salpêtrière </func> </pers.ind>
Der Bürgermeister Delanoë Der <pers.ind> <func>Bürgermeister</func> <name> Delanoë </name>	*Le maire Delanoë* Der <pers.ind> <func>maire</func> <name> Delanoë </name>

</pers.ind>	</pers.ind>
Bertrand Delanoë, der Bürgermeister von Paris <pers.ind> <name> Bertrand Delanoë </name>, der <func> Bürgermeister von <loc.adm.town> Paris </loc.adm.town> </func> </pers.ind>	*Bertrand Delanoë, le maire de Paris* <pers.ind> <name>Bertrand Delanoë </name>, le <func>maire de <loc.adm.town> Paris </loc.adm.town> </func> </pers.ind>
Herr Martin, der türkische Botschafter in Frankreich <pers.ind> <title> Herr </title> <name> Martin </name>, der <func> türkische Botschafter in <loc.adm.nat> Frankreich </loc.adm.nat> </func> </pers.ind>	*Monsieur Martin, l'ambassadeur de Turquie en France* <pers.ind> <title> Monsieur </title> <name> Martin </name>, l' <func> ambassadeur de <loc.adm.nat> Turquie </loc.adm.nat> en <loc.adm.nat> France /loc.adm.nat> </func>
General De Gaulle <pers.ind> <title> General </title> <name> De Gaulle </name> </pers.ind>	*le général De Gaulle* Le <pers.ind> <title> Général </title> <name> De Gaulle </name> </pers.ind>
qualifier	
Der konservative Christoph Blocher Der <pers.ind> <qualifier> konservative </qualifier>\ <name> Christoph Blocher </name> </pers.ind>	*Le socialiste Bertrand Delanoë* Le <pers.ind> <qualifier> socialiste </qualifier>\ <name> Bertrand Delanoë </name> </pers.ind>
name	
von Lange <pers.ind> <name> von Lange </name> </pers.ind>	*De Gaulle* <pers.ind> <name> De Gaulle </name> </pers.ind>
demonym	
Der Engländer Tony Blair erklärt…. Der <pers.ind> <demonym> Engländer </demonym> <name>Tony Blair</name> </pers.ind>	*L'anglais Tony Blair a déclaré….* L' <pers.ind> <demonym> anglais </demonym> <name>Tony Blair</name> </pers.ind>

D. Tricky cases for Person

- Annotation of particle: The (noble) particle is part of the name when it is present:

  M. le comte de Metternich en a aussitôt informé la députation

  ```
  <pers.ind>
        <comp.title>M.</comp.title> le
        <comp.title>comte</comp.title>
        <comp.name>de Metternich</comp.name>
  </pers.ind>
  ```

- Hesitation between title and function:
  'Titles' correspond to expressions that indicate a social status and can be used to address to a person, especially orally (Mr., Mlle, Lord, *etc.*).
  'Functions' correspond to expressions that indicate the profession of a person.
  The distinction is not always easy, as certain expressions can be used as both honorific titles and roles or professions, especially within the military domain.
  We apply the following rule(s):
  - military titles/grades are always 'title'
  - when hesitating, try to see if the word could be used to address a person orally:
    "Bonjour Monsieur Martin", "Bonjour Général Martin" => valid utterance, therefore annotation as `comp.title`
    "Bonjour footballeur Martin" => less valid utterance, therefore annotation of footballeur as `comp.func`

- If the title is split in 2 parts

  der Baron Jakobi-Klöst zu Erfurt
  The correct title is *Baron zu Erfurt* but we do not deal with split components. In this case, the annotation should be:

  ```
  <pers.ind>
        <comp.title>Baron</comp.title>
        <comp.name>Jokobi-Klöst</comp.name> zu
        <loc.adm.town>Erfurt</loc.adm.town>
  </pers.ind>
  ```

3.2 Organisations

A. Subtypes

- `<org.ent>`: A company which sells products or provides services that are not only administrative. It includes both private and public companies, as well as hospitals, schools, universities, political parties, trade unions, police, gendarmerie, churches, (named) armies, sportive clubs, etc. Organizations of administrative nature mainly are excluded.

Die Peugeot Gesellschaft

```
Die <org.ent>
 <name> Peugeot </name>
     Gesellschaft
</org.ent>
```

Ich arbeite bei Peugeot

```
Ich arbeite bei
<org.ent> Peugeot
</org.ent>
```

Die UNESCO

```
Die <org.ent> UNESCO
</org.ent>
```

Die Rote Armee

```
Das <org.ent>
     Rote Armee
</org.ent>
```

Die Grüne Partei: 'Partei' is part of the name of this party (GPS)

```
Die <org.ent> Grüne Partei
</org.ent>
```

Die Partei JungsozialistInnen Schweiz: 'Partei' is not part of the name of this party (juso)

```
Die <org.ent> Partei
<name> JungsozialistInnen Schweiz
</name>
</org.ent>
```

Die Gewerkschaft UNIA

```
die <org.ent>
     Gewerkschaft
    <name> UNIA </name>
</org.ent>
```

Die Gewerkschaft des Verkehrspersonals

```
Die <org.ent> Gewerkschaft des
Verkehrspersonals </org.ent>
```

La société Peugeot

```
La <org.ent> société
    <name> Peugeot </name>
</org.ent>
```

Je travaille chez Peugeot

```
Je travaille chez
<org.ent> Peugeot
</org.ent>
```

L' UNESCO

```
L' <org.ent> UNESCO
</org.ent>
```

L'Armée Rouge

```
L' <org.ent> Armée
    <name> Rouge </name>
</org.ent>
```

l'hôpital d'instruction des armées du Val-de-Grâce

```
L' <org.ent>
     hôpital d'instruction des
armées
    du
    <name> Val-de-Grâce </name>
</org.ent>
```

La parti socialiste: 'parti' is part of the name of this party (PS)

```
La <org.ent> parti socialiste
</org.ent>
```

La parti Europe Écologie: 'parti' is not part of the name of this party (EE)

```
La  <org.ent> parti
<name> Europe Écologie </name>
</org.ent>
```

Le syndicat FSU

```
Le <org.ent> syndicat
    <name> FSU </name>
</org.ent>
```

Le syndicat national de la magistrature

```
Le <org.ent> syndicat national
de la magistrature
</org.ent>
```

- <org.adm> refers to an organisation which plays a mainly administrative role. It is often an administrative and/or geographical division. This includes town halls, city council, regional council, state council, federal council, named government, minister, parliament, prefectures, ministries, dioceses, tribunal, court, government treasury, public treasury, international org.

Die Stadtverwaltung Bern
```
Die <org.adm>
Stadtverwaltung
<loc.adm.town> Bern
</loc.adm.town>
</org.adm>
```

La Mairie de Paris
```
La <org.adm> mairie de
    <loc.adm.town> Paris
</loc.adm.town>
</org.adm>
```

Das Bistum Basel
```
Das <org.adm>
     Bistum
    <loc.adm.town> Basel
</loc.adm.town>
</org.adm>
```

Le diocèse de Blois
```
Le <org.adm>
     diocèse  de
    <loc.adm.town> Blois
</loc.adm.town>
</org.adm>
```

- <org.ent.pressagency>: A specific subtype used for newspapers material (*AFP*, *Reuters*)

3.3 Locations

A. Administrative locations: loc.adm.*

<loc.adm.*> refers to a territory with a geopolitical border. The subtypes of <loc.adm> correspond to different granularities of territory chunks.

- district, city: <loc.adm.town> includes cities and all smaller units:
  - city, village, hamlet, locality, commune;
  - part of the city: district, borough, etc.

Zürich
```
<loc.adm.town> Zürich
</loc.adm.town>
```

Der Kreis 4
```
Der <loc.adm.town>Kreis 4
</loc.adm.town>
```

Paris
```
 <loc.adm.town> Paris
</loc.adm.town>
```

Die Stadt Zürich
```
Die <loc.adm.town> Stadt
Zürich</loc.adm.town>
```

La Bollline

```
<loc.adm.town> La Bolline
</loc.adm.town>
```

Val de Crüye

```
<loc.adm.town> Val de Crüye
</loc.adm.town>
```

Maison Blanche [7]

```
<loc.adm.town> Maison Blanche
</loc.adm.town>
```

La ville de Paris

```
La <loc.adm.town> ville de
Paris </loc.adm.town>
```

Big Apple

```
<loc.adm.town> Big
Apple</loc.adm.town>
```

Le 13e arrondissement

```
Le <loc.adm.town> 13e
arrondissement </loc.adm.town>
```

La ville rose

```
La <loc.adm.town> ville rose
</loc.adm.town>
```

- region: `<loc.adm.reg>` refers to internal divisions within a state[8] and includes all units between country and city levels: administrative and traditional regions, departments, counties, departmental districts, Swiss cantons, including the associated municipalities communities of municipalities, urban communities, etc.

Die Autonome Gemeinschaft Baskenland

```
Die <loc.adm.reg> Autonome
Gemeinschaft Baskenland
</loc.adm.reg>
```

Im Süden von Israel

```
<loc.adm.reg>
    <qualifier> Im Süden
</qualifier>
    von <name>
   <loc.adm.nat> Israel
</loc.adm.nat>
    </name>
</loc.adm.reg>
```

la CAPS

```
la <loc.adm.reg> CAPS
</loc.adm.reg>
```

Au sud d'Israël

```
au <loc.adm.reg>
sud d'
   <loc.adm.nat> <name>Israël
</name>
</loc.adm.nat>
   </loc.adm.reg>
```

Le Pays basque espagnol

```
Le <loc.adm.reg> Pays basque
espagnol </loc.adm.reg>
```

- national: `<loc.adm.nat>` for countries.

Die Schweiz, Vereinigtes Königreich, die Vereinigten Staaten, Andorra; Monaco, la France, le Royaume-Uni, les États-Unis.

Das Vereinigte Königreich *Le Royaume-Uni*

[7] *Maison Blanche* is a district of Paris.

[8] Monaco is either a state or a city, but not a region. The Gaza Strip is a region of Israel and is therefore annotated with `<loc.adm.reg>`.

```
Das <loc.adm.nat> Vereinigte
Königreich </loc.adm.nat>
```

```
Le <loc.adm.nat> Royaume-Uni
</loc.adm.nat>
```

- supranational: `<loc.adm.sup>` refers to world regions, continents, etc. :

  Der Nahe Osten, das Baskenland, Katalonien, der Commonwealth, der Norden, le Moyen Orient;
  le Pays basque, la Catalogne, le Commonwealth, l'Afrique subsaharienne, le Sud [9]

Das Baskenland

```
Das <loc.adm.sup> Baskenland
</loc.adm.sup>
```

Le Pays basque

```
Le <loc.adm.sup> Pays basque
</loc.adm.sup>
```

Die Region um den Atlas

```
Die <loc.adm.sup>
     Region
    um den
    <name> Atlas </name>
</loc.adm.sup>
```

La région de l'Atlas

```
La <loc.adm.sup>
     région
    de l'
    <name> Atlas </name>
</loc.adm.sup>
```

About hesitation between loc.adm.reg and loc.adm.nat: overloaded definition of loc.adm.reg

The concept of 'country' as a stable state entity with clear borders is not well adapted for historical corpora. Geopolitical entities naturally evolved through time and what was a Principality at some time might have become a state. In some cases it might therefore be difficult to choose between the tags `loc.adm.nat` and `loc.adm.reg` for e.g. *Bavière, Dalmatie*.

We decided to overload the definition of loc.adm.reg, as a tag which can be used for

1) internal divisions within a states (= contemporary regions)
2) for borderline cases where it is difficult to choose between nat and reg (= fuzzy reg)

Consequently, entities which used to be a kind of independent states or strong entities but were not real states, or are not anymore today, are annotated as loc.adm.reg[10].

B. Physical places: loc.phys.*

- terrestrial physical locations: `loc.phys.geo`

[9] In the sense of the countries of the South. In other contexts, the south could designate other geographical locations (*le Sud de la France*).
[10] Our intention here is not to hide historical realities, but to adopt a pragmatic position w.r.t our NE objectives.

Geonyms[11] include names given to natural geographical spaces, such as deserts, mountains, mountain chains, glaciers, plains, chasms, plateaus, valleys, volcanoes, canyons, etc.

Der Ätna
```
Der <loc.phys.geo> Ätna
</loc.phys.geo>
```

L'Etna
```
L' <loc.phys.geo> Etna
</loc.phys.geo>
```

Die Wüste Gobi
```
Die <loc.phys.geo>
Wüste <name> Gobi </name>
</loc.phys.geo>
```

Le desert de Gobi
```
Le <loc.phys.geo>
désert de <name> Gobi </name>
</loc.phys.geo>
```

- aquatic physical sites: `loc.phys.hydro`
  Hydronyms [12] refer to water bodies [13], such as rivers, streams, ponds, marshes, lakes, seas, oceans, marine currents, canals, springs, etc.

Die Spree
```
Die <loc.phys.hydro> Spree
</loc.phys.hydro>
```

La Seine
```
La <loc.phys.hydro> Seine
</loc.phys.hydro>
```

Der Canal Saint-Martin
```
Der <loc.phys.hydro>
     Canal
    <name> Saint-Martin </name>
</loc.phys.hydro>
```

Le Canal Saint-Martin
```
Le <loc.phys.hydro>
     Canal
    <name> Saint-Martin </name>
</loc.phys.hydro>
```

- astronomical physical places: `loc.phys.astro` includes planets, stars, galaxies, etc., and their parts.

Der Mond
```
Der <loc.phys.astro> Mond
</loc.phys.astro>
```

La Lune
```
La <loc.phys.astro> Lune
</loc.phys.astro>
```

Die Milchstrasse
```
Die <loc.phys.astro>
Milchstrasse </loc.phys.astro>
```

la mer de la tranquillité
```
La <loc.phys.astro> mer de la
    <name> tranquillité </name>
</loc.phys.astro>
```

C. Pathways: loc.oro

Oronyms (`loc.oro`) are streets, squares, roads, highways, etc. They cannot be considered physical or administrative places. They are often artificial but sometimes natural, and are not necessarily geopolitical: this justifies not classifying them as `loc.adm` and `loc.phys`.

[11] Definition taken from Mickaël Tran's thesis, Université de Tours, 2006, p. 84
[12] Definition taken from Mickaël Tran's thesis, Université de Tours, 2006, p. 84
[13] We include water streams as well.

Die Autobahn A6

```
Die <loc.oro>
Autobahn <name> A6 </name>
</loc.oro>
```

Die A6

```
Die
<loc.oro> A6
</loc.oro>
```

Die Nordring Autobahn

```
Die <loc.oro>
  <name> Nordring </name>
 Autobahn
</loc.oro>
```

Der Nordring

```
Der <loc.oro>
    <name> Nordring </name>
</loc.oro>
```

place de l'Abbé Georges Hénocque

```
<loc.oro> place de l'
<name>
  <pers.ind>
  Abbé Georges Hénocque
</pers.ind>
</name>
</loc.oro>
```

rue de Vaugirard (Vaugirard is a village)

```
<loc.oro> rue de
<name> Vaugirard </name>
</loc.oro>
```

la 118

```
la <loc.oro> 118
</loc.oro>
```

le triangle de Rocquencourt

```
le  <loc.oro>
triangle de <name>
    <loc.adm.town> Rocquencourt
</loc.adm.town>
    </name>
</loc.oro>
```

L'autoroute A6

```
L' <loc.oro> autoroute
    <name> A6 </name>
</loc.oro>
```

rue des Glycines

```
<loc.oro>
rue des
<name> Glycines </name>
</loc.oro>
```

D. Buildings : loc.fac

Named buildings (train station, museum, ..) as well as their extensions (stadium, campus, university, camping…) are annotated. `loc.fac` often refers to the physical location of an organisation.

Zürich Hauptbahnhof

```
<loc.fac>
      <name> Zürich </name>
       Hauptbahnhof
</loc.fac>
```

Bern Bümpliz Nord

La gare de Rungis

```
La <loc.fac>
gare  de
<name> Rungis </name>
</loc.fac>
```

la gare Saint-Germain Grande Ceinture

```
la <loc.fac> gare
<name> Saint-Germain Grande
```

```
<loc.fac>
<name> Bern Bümpliz Nord </name>
</loc.fac>
```

Der ehemalige Bahnhof Letten

```
Der <loc.fac>
      ehemalige Bahnhof
      <name> Letten </name>
</loc.fac>
```

Schloss Kyburg

```
<loc.fac>
     Schloss
    <name> Kyburg </name>
</loc.fac>
```

Die Kyburg

```
Die <loc.fac>
    <name> Kyburg </name>
</loc.fac>
```

```
Ceinture </name>
</loc.fac>
```

l'ancienne gare de Rungis

```
l' <loc.fac>
ancienne gare de
<name> Rungis </name>
</loc.fac>
```

le palais de l'Élysée

```
Le <loc.fac>
    palais de l'
    <name> Élysée </name>
</loc.fac>
```

l'Élysée

```
l' <loc.fac>
    <name> Élysée </name>
</loc.fac>
```

E. Addresses: loc.add.*

- physical addresses: `loc.add.physic`
  As opposed to a `loc.oro`, an address is a point in space (e.g. a point in a street)

Ich wohne in der Shilstrasse 15 3. Stock

```
Ich wohne in der
<loc.add.phys>
    <loc.oro>
        <name> Sihlstrasse </name>
    </loc.oro>
    <address-number> 15 </address-number>
    <other-address-component> 3. Stock </other-address-component>
</loc.add.phys>
```

9 place de Rungis

```
<loc.add.phys>
    <address-number> 9 </address-number>
<loc.oro>
 place de
<name><loc.adm.town> Rungis </loc.adm.town></name>
</loc.oro>
</loc.add.phys>
```

J'habite 15 rue de Vaugirard escalier 2

```
J' habite
<loc.add.phys>
    <address-number> 15 </address-number>
    <loc.oro>
         rue de <name> Vaugirard </name>
    </loc.oro>
```

```
    <other-address-component> escalier 2 </other-address-component>
</loc.add.phys>
```

- electronic addresses: `loc.add.elec`
  Electronic coordinates: a telephone or fax number, url, E-Mail address, frequency radio, social network identifiers (*Facebook, Twitter*) or tools for internet communication (*Skype*), etc.

Meine Nummer lautet 01 69 85 80 02

```
Meine Nummer lautet
<loc.add.elec>
<name> 01 69 85 80 02 </name>
</loc.add.elec>
```

mon numéro est le 01 69 85 80 02

```
mon numéro est le
<loc.add.elec>
 <name> 01 69 85 80 02
</name>
</loc.add.elec>
```

Mein Skype-Name ist jean.dupont

```
Mein Skype-Name ist
<loc.add.elec>
<name> jean.dupont </name>
</loc.add.elec>
```

mon identifiant skype est jean.dupont

```
mon identifiant skype est
<loc.add.elec>
<name> jean.dupont </name>
</loc.add.elec>
```

Radio Bleue auf 98.8 MHz

```
<prod.media> Radio Bleue
</prod.media> auf
<loc.add.elec>
      <name> 98.8 MHz </name>
</loc.add.elec>
```

Radio Bleue sur 98.8 MHz

```
<prod.media> Radio Bleue
</prod.media> sur
<loc.add.elec>
      <name> 98.8 MHz </name>
</loc.add.elec>
```

Folgt mir auf Twitter unter @leguidedannotation

```
Folgt mir auf
<prod.soft>
      <name> Twitter </name>
</prod.soft> unter
<loc.add.elec>
      <name> \@leguidedannotation
</name>
</loc.add.elec>
```

suivez-moi sur Twitter à @leguidedannotation

```
suivez-moi sur
<prod.soft>
      <name> Twitter </name>
</prod.soft> à
<loc.add.elec>
      <name>
\@leguidedannotation </name>
</loc.add.elec>
```

Special cases for websites:

- reference to the access of the website: `<loc.add.elec>` :
  Lesen sie den Artikel auf lemonde.fr;
  retrouvez cet article sur lemonde.fr

- reference to the website as a whole: <prod.media> :
  Interview auf lemonde.fr, mediapart.fr zeigt, dass Eric Woerth 50.000 Euro erhalten hat; Interview à retrouver sur lemonde.fr, mediapart.fr indique que Eric Woerth a bien touché 50.000 euros

- reference to the company that publishes the site: <org.ent> :
  Sarkozy bemängelt mediapart.fr;
  Sarkozy dénonce mediapart.fr

Site addresses (www.radio-france.fr) are annotated as <loc.add.elec>. However *Le site internet Radio France* is not a entity named in itself (we annotate only *Radio France* with prod.media).

3.4 Human Productions

A. Media production: prod.media

Anything that is broadcast in the press, on radio or television: newspapers, magazines, broadcasts, sales catalogues, etc.

Media products are excluded: films, TV films, ets.

Mickaël Vendetta möchte La Ferme des Célébrités verlassen

```
<pers.ind>
<name> Mickaël Vendetta </name>
</pers.ind>
möchte
<prod.media>
      <name> La Ferme des
Célébrités </name>
</prod.media> verlassen
```

Mickaël Vendetta veut quitter La Ferme des Célébrités

```
<pers.ind>
<name>MickaëlVendetta</name>
</pers.ind>
veut quitter
<prod.media>
      <name>La Ferme des
Célébrités</name>
</prod.media>
```

Le Monde hat einen Artikel über 1984 von George Orwell.

```
<prod.media>
      <name> Le Monde </name>
</prod.media>
aat einen Artikel über
<prod.art>
      <name>1984</name>
</prod.art>
von
<pers.ind>
  <name> George Orwell </name>
</pers.ind>
```

Le Monde a publié un article sur 1984 de George Orwell.

```
<prod.media>
      <name>Le Monde</name>
</prod.media>
a publié un article sur
<prod.art>
      <name>1984</name>
</prod.art> de
<pers.ind>
      <name> George Orwell</name>
</pers.ind>
```

Ambiguity between prod.media and org.ent:
Sometimes the name of a media (`prod.media`) is also the name of a company (`org.ent`): *France Inter, le Monde, etc*. The following test can be applied to distinguish between the two cases: we insert *das Medium/das Unternehmen; le média/la société* before the expression to be annotated.

- [das Medium] *Le Monde veröffentlicht einen Artikel*
  (not [das Unternehmen] *Le Monde veröffentlicht einen Artikel*) => `<prod.media>`

- *Willkommen bei* [dem Medium] *France Inter*
  (not *Willkommen bei* [dem Unternehmen] *France Inter*) => `<prod.media>`

- [das Unternehmen] *Libération feuerte drei Journalisten*
  (not [das Medium] *Libération feuerte drei Journalisten*) => `<org.ent>`

- [das Unternehmen] *Le Monde ist ein Opfer von politischer Einmischung*
  (not [das Medium] *Le Monde ist ein Opfer von politischer Einmischung*) => `<org.ent>`

- [le média] *Le Monde publie un article*
  (not [la société] *Le Monde publie un article*) => `<prod.media>`

- *Bienvenue sur* [le média] *France Inter*
  (not *Bienvenue sur* [la société] *France Inter*) => `<prod.media>`

- [la société] *Libération a licencié trois journalistes*
  (not [le média] *Libération a licencié trois journalistes*) => `<org.ent>`

- [la société] *Le Monde est victime d'une ingérence politique*
  (not [le média] L*e Monde est victime d'une ingérence politique*) => `<org.ent>`

B. Doctrine: prod.doctr
The label applies to socialism, communism, Buddhism, Protestantism...

Der Sozialismus kann als .. eingeschätzt werden

```
Der <prod.doctr>
    <name> Sozialismus </name>
</prod.doctr>
```

Le socialisme est un un type d'organisation sociale...

```
Le <prod.doctr>
    <name>socialisme</name>
</prod.doctr>
```

3.5 Time

Reminder : relative or subordinate phrases are not part of an entity.

A. Date : time.date

In impresso we annotate **absolute dates only**: `time.date.abs`, without components.

An absolute date is a date whose position on the calendar can be deduced by the sole information present in the date (or temporal expression), without any context.

These are absolute dates:

Montag 25. Januar 2010
```
<time.date.abs>
Montag 25. Januar 2010
</time.date.abs>
```

lundi 25 janvier 2010
```
<time.date.abs>
lundi 25 janvier 2010
</time.date.abs>
```

However, as soon as an explicit marker of relativity (for example *nächster; prochain*) is specified (*Meeting am nächsten Dienstag; rendez-vous mardi prochain*), we have a relative date and do not annotate it. Determiners (*der, die, das; le, la, les, l'*) are only included in the entity if it has a function equivalent to a preposition (*à, en*).

Um 300 vor Christuns
```
<time.date.abs>
um 300 vor Christus
</time.date.abs>
```

en 300 avant JC
```
<time.date.abs>
en 300 avant JC
</time.date.abs>
```

Das Jahr 1990
```
Das <time.date.abs>
Jahr 1900
</time.date.abs>
```

l'année 1900
```
L' <time.date.abs>
année 1900
</time.date.abs>
```

Das 21. Jahrhundert
```
Das <time.date.abs>
    21. Jahrhundert
</time.date.abs>
```

le 21e siècle
```
Le <time.date.abs>
21e siècle
</time.date.abs>
```

Das dritte Jahrtausend
```
Das <time.date.abs>
    <millenium> dritte Jahrtausend
</millenium>
</time.date.abs>
```

le troisième millénaire
```
le <time.date.abs>
    troisième millénaire
</time.date.abs>
```

No annotation for *ère professionelle* in :

Die fünfte Spielerin der professionellen Ära.

La cinquième joueuse de l'ère professionnelle.

3.6 Non-annotated entities

- Names of diseases (*AIDS, Grippe A; SIDA,* etc.)
- Psychological phenomena (*Ödipuskomplex; syndrome de Stockholm*, etc.)
- Scientific terms cannot be reduced to a product (*DNA, ADN,* etc.)
- Teaching programmes (*Staps, DEUG*, etc.)

- Special contracts (*le contrat Coca-Cola/Danone,* etc.)
  However: in *le contrat Coca-Cola*, the entity *Coca-Cola* is annotated (`org.ent`).
- Political and/or judicial matters (*Watergate, Monica-gate; affaire Dickinson,* etc.).
  Optional: these may fall into a category depending on the assessments of the annotators.
- Climatic phenomena (*der Sturm Yinthya, le Mistral,* etc.).
  Optional: these may fall into a category depending on the assessments of the annotators.
- Social phenomena (*l'immigration arménienne* [14], etc.).
  Optional: these may fall into a category depending on the assessments of the annotators.

  NOTE: In some cases, it is still necessary to annotate the components of these expressions.
  - we do not annotate *Stockholm Syndrome* but we must annotate *Stockholm* (`<loc.adm.town>`)
  - we do not annotate *complex d'Oedipus* but we must annotate *Oedipus* (`<pers.ind>`)

4. Entity linking

- Entity mentions are linked against Wikidata. We link the following entity types: `pers.ind`, `org.*`, `loc.*` and `prod.media`.

- **Entity components** and **nested entities** are excluded from the linking.
  => In the case of a long entity, the full string needs to be annotated: e.g. in *le chancelier d'Empire Ebert,* the full string (pers.ind) is linked, not only *Ebert* (comp.name)
  => Even if tempting, do not annotate nested entities; e.g. *H.C. Lausanne* (Lausanne hockey club), only the full string *H.C. Lausanne* is linked, not *Lausanne.*

- In case the entity to which the mention refers to is **not present in Wikidata**, the mention is marked as referring to a **NIL** entity.

- If a mention has a **metonymic sense** on top of the **literal sense**, **both senses** (i.e, both annotations) need to be linked.

  We distinguish between *weak* and *strong* metonymy, where:
  - *weak metonymy* corresponds to the case where the literal and metonymic senses are quite intermingled and cannot be easily set apart. This is the case with countries names, often used to refer to the country's government (*France signed a treaty with Germany last week*). In such cases, both geographical and political sense are present to the mind, and these entities are often referred to as GeoPolitical Entities (GPE).
    In *impresso*, such entities have 2 annotations: loc.adm.nat for the literal sense, and org.ent for the metonymic sense.
    For entity linking of these country-to-org cases, although Wikidata contains some

[14] However, this term is annotated if it refers to a group of people rather than a process, see section 2.3.1.2.

government entities, *impresso* rule is to link both mention annotations with the country Wikidata entity, which refers to both geographical and political aspects (e.g. the description of France “republic with mainland in Europe and numerous oversea territories”, and of Germany “federal parliamentary republic in central-western Europe”).

- *strong metonymy* corresponds to the case where the literal and metonymic senses have more distinct traits and can less easily be used jointly. This is the case with city names, often used to refer to sportive team, or also to governments. In such cases, it is more difficult to fully map the geographical sense to the metonymic one.
  In such cases, the *impresso* rule is to link to the literal entity for the literal sense and, when possible, to the metonymic entity for the metonymic sense.

In short, the rules are as follows:

- literal sense of a country mention => literal wikidata entity
- metonymic sense of a country mention => literal wikidata entity
- literal sense of any other mention => literal wikidata entity
- metonymic sense of any other mention => metonymic wikidata entity

Since metonymy annotation and linking is a difficult task and is not the main focus of the HIPE evaluation lab, our goal is not to cover all specific cases, but to capture these phenomena at a high level and in the most meaningful way.

- For **historical entities** for which it is difficult to determine the exact corresponding referent, or for which there is no exact historical referent but a contemporary one in Wikidata, one has to pick the contemporary entity (for example the *Serbia* of the 19th century is not the same as the contemporary Serbia referred to in Wikidata, but in the absence of historical entity one take the contemporary Serbia Wikidata entry).

5. Quick guide and concrete considerations

6.1 Hesitations

A. Checking

If you need to double check a point, please use these resources:

- for German, Duden: http://duden.de
- for French, Larousse (tab ‘Dictionary’ or ‘Encyclopedia’): https://www.larousse.fr/dictionnaires/francais

In case you suspect something to be a named entity but a quick check on the above mentioned resources and/or Wikipedia does not give information, skip the annotation.

B. Reporting hesitations

For any dubious cases, please report your questions with screenshot and comments in a dedicated file.

C. Inception mini-tutorial
https://docs.google.com/document/d/1jtfwQlvyg8YFJ4SR2xeNURrVGI7wLjHD3tgwzIbmfmk/edit?usp=sharing

6.2 Overview of types, subtypes and components

Entity types and subtypes	
`pers.ind`	A single person (*Roger Federer*)
`pers.ind.articleauthor`	A single person who is the author of an article.
`pers.coll`	A named group of people including musical groups (die *Beatles*, La *Mano Negra*). (note: die *Schweizer*, Les *français* are **not** annotated.)
`org.ent`	Organization that markets products or provides services (Die *Peugeot Gesellschaft,* Die *Waid;* La *société Peugeot*, la *Pitié-Salpêtrière*). (note: Die *schweizer* Polizei; la police francaise ist **not** annotated)
`org.ent.pressagency`	Special type related to newspaper to spot press agencies.
`org.adm`	Organization that plays a mainly administrative role (Die *Stadtverwaltung Bern*; la *mairie de Paris*). (note: Das *Departement für auswärtige Angelegenheiten*; Le *Ministère des Affaires Étrangères* is **not** annotated)
`loc.adm.town`	District, locality, hamlet, village, city, etc. (*Paris*, *Val de Crüye*).
`loc.adm.reg`	Cantons, communities of municipalities, departments, regions, etc. (*Autonome Gemeinschaft Baskenland;* les *Bouches du Rhône*, Le *Pays-Basque espagnol*).
`loc.adm.nat`	Countries (*Schweiz*; *France*).
`loc.adm.sup`	World regions, continent (*Maghreb; Pays-Basque*).
`loc.phys.geo`	Mountains, plains, plateaus, caves, volcanoes, canyons (Die *Alpen,* Der *Vesuv; gouffre de Padirac*, Le *mont Ventoux*).

loc.phys.hydro	Oceans, seas, rivers, streams, ponds, marshes (Der *Atlantik,* Der *Golfstrom;* La *Seine*, Le *Lac Paladru*).
loc.phys.astro	Planets, stars, galaxies and their parts (Der *Mond*, Die *Milchstrasse;* La *terre,* la *mer de la Tranquillité*).
loc.oro	Refers to roads, highways, streets, avenues, squares, etc. (Die *Autobahn A6;* L'*autoroute A6*).
loc.fac	Refers to the buildings (Der *Prime Tower;* Le *Palais de l'Élysée*).
loc.add.phys	Refers to physical addresses (*LIMSI-CNRS*, *Bâtiment 508, BP133, 91403 Orsay Cedex*).
loc.add.elec	Refers to electronic contact information (telephone and fax numbers, URL, e-mail address, identification of social network or Internet communication tools, etc., *http://www.limsi.fr/, 01-69-85-80-00*)
Loc.unk	Type used when it is not possible to choose among other location types.
prod.media	Newspapers, magazines, broadcasts, sales catalogues, etc. (*Die Zeit;* Le *Figaro*, *Le sept à huit, La ferme célébrités*).
prod.doctr	Political, philosophical, religious, sectarian doctrines. (Der *Sozialismus, Theravada Buddhismus; Zeugen Jehovas;* Le *socialism*, le *bouddhisme theravâda*,le *structuralism*, la *scientology*).
time.date.abs	An absolute date (*Sonntag der 13. November 2016; lundi 25 janvier 2010*)
Component	
name	is the only transversal component and is applied to any class except time. (Die *Peugeot* Gesellschaft; la société *Peugeot*)
comp.name	The component includes first, middle and last names as well as nickname and initials of individuals (*Samuel L. Jackson, S.L.J.*)
comp.title	Title or designator of a person. (*Herr* Chirac, *Ihre Hoheit* Rainier*; M.* Chirac, *Son Altesse* le prince Rainier).
comp.qualifier	A qualifier specifies a person in the form of a qualifying adjective. (Der *konservative* Christoph Blocher; le *socialiste* Bertrand Delanoë)
comp.function	A function or job of a named person. (*Bürgermeister* Ann Hidalgo von Paris; *maire* de Paris Anne Hidalgo).

`comp.demonym`	The geographical origin of a person (Le *français* Alain Vigneron).

Acknowledgements

This document is based on the QUAERO guidelines developed by Sophie Rosset, Cyril Grouin and Pierre Zweigenbaum. We warmly thank them for having shared with us their experience and the QUAERO guidelines document sources.