

NewsEye

Named Entity and Stance Annotation Guidelines

Version: 3.1 - March 2020

Initially based on version 2.0 of the Impresso NE annotation guidelines¹

[1. Preamble](#)

[2. General instructions](#)

[2.1 Entity types and subtypes](#)

[2.2 Named entity mention lexical characteristics](#)

[2.3 Nesting and special constructions](#)

[2.4 Ambiguities](#)

[3. Entities](#)

[3.1 Person](#)

[3.2 Organisations](#)

[3.3 Locations](#)

[3.4 Human production](#)

[3.5 Non-annotated entities](#)

[4. Quick guide and concrete considerations](#)

[4.1 Hesitations](#)

[4.2 Overview of types, subtypes and components](#)

[5. Stance annotation guidelines](#)

[6. Named entity linking guidelines](#)

[6.1 How Specific Should Linked Entities Be?](#)

[6.2 Metonymy](#)

[6.3 Can Mention Boundaries Overlap?](#)

[ANNEX A Main changes w.r.t Quaero v1](#)

[ANNEX B Main changes w.r.t Impresso v2](#)

[ANNEX C Main changes w.r.t NewsEye v3](#)

¹ By Maud Ehrmann, Camille Watter, Matteo Romanello, Simon Clematide (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

While the part of the guidelines on stance detection annotation is new, the NewsEye NE annotation guidelines are derived from *Impresso* NE annotation guidelines which are derived from Quaero guidelines². 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 corpora³. *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. Followingly, the NewsEye guidelines are intended to be compatible with the *Impresso* ones, in order to allow the produced datasets to be compatible too, and so that both projects (and the community at large) can benefit of combined efforts and a significant amount of compatible training data, rather than from independent and incompatible smaller collections.

Application context

The objective is to extract information from historical newspaper articles, in view of supporting the search, filtering and analysis of large collections 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).

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:

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_lqgcHaa3S-nZ2Sr71iZ5YbecDCc/edit#

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

Type	Subtypes
pers	pers.articleauthor
loc	
org	
(prod)	prod.media

In the NewsEye NE annotation, the subtypes are very limited:

- <pers.articleauthor> is a specific subtype of person describing the author of an article, especially useful for newspapers. Every other type of person NE should be simply annotated with <pers>. This is further detailed in Section 3.1
- <prod.media> is a specific subtype of human production described in Section 3.4. As this is the only type of human production we wish to annotate in NewsEye, we will never actually use the <prod> annotation but only the <prod.media> annotation
- <org> and <loc> are for organisations and locations. No subtypes are to be taken into account.

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 name⁴. 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 names⁵, 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 discussion⁶.

Phrases such as

- *Die präkolumbianische Zivilisation, la civilisation précolombienne*
- *l’armée bavaroise*

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”.

- *les forces tchadiennes*
- *le gouvernement français*

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

Phrases such as:

- *le gouvernement Franco*
le <org> gouvernement
 <pers> Franco </pers>
 </org>

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.
- 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...
- 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 baloise

=> no annotation

Zürichputsch

=> French translation (decomposition): le putsh de Zurich (Putsch von Zürich)

=> annotation of “Zürich”

`<loc>Zürich</loc>putsch`

Donaufestungen

=> Festungen an der Donau

=> annotation of “Donau”

`<loc>Donau</loc>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>ägyptische Ministerpräsident Al-Nokraschi</pers>`

The connecting “s” in German compounds is *not* annotated:

Völkerbundsmitgliedern

=> only *Völkerbund* is annotated

`<org>Völkerbund</org>smitgliedern`

2.3 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, with a limit of nested entities of depth 1, i.e. a nested entity cannot contain a nested entity (note that entity linking and stances are not concerned by nested entities).

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

La société du Parc du Creux-du-Vent...
 <org>société du
 <loc>Parc du Creux-du-Vent</loc>
 </org>

Le maire de Paris Bertrand Delanoë a déclaré
 <pers>
 maire de <loc>Paris </loc> Bertrand Delanoë
 </pers>

dem Preussischen Staatsminister der auswärtigen Angelegenheiten, Graf von Goltz
 <pers>
 Preussischen
 Staatsminister der auswärtigen Angelegenheiten
 Grafvon Goltz
 </pers>

- 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> Bodensee </loc>,
 <loc> Starnberger See </loc>, und
 <loc> Müritz </loc>

vallées de la Lorraine, de l'Alsace et de la Champagne
 <loc> vallées de la Lorraine</loc>,
 <loc> de l'Alsace </loc> et
 <loc> de la Champagne </loc>

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>`Madame Chirac `</pers>` ...

Ost und Mitteleuropa...

Ost und `<loc>`Mitteleuropa`</loc>`....

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>` Shekau,
chef de l'une des trois factions de
`<org>`Boko Haram`</org>` et
fondateur historique du groupe
`</pers>`

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>` DAISY `</org>` das

`<org>` Dynamische Auskunfts- und Informationssystem `</org>`

Agipi association d'assurés pour la prévoyance, la dépendance et l'épargne-retraite

`<org>` Agipi`</org>`\

`<org>` Association d'assurés pour la prévoyance , la dépendance et l'épargne-retraite

`</org>`

D. Difficult example(s)

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

`<pers>`

bekannte

Irländer

Theobald Wolfe Tone

`</pers>`

2.4 Ambiguities

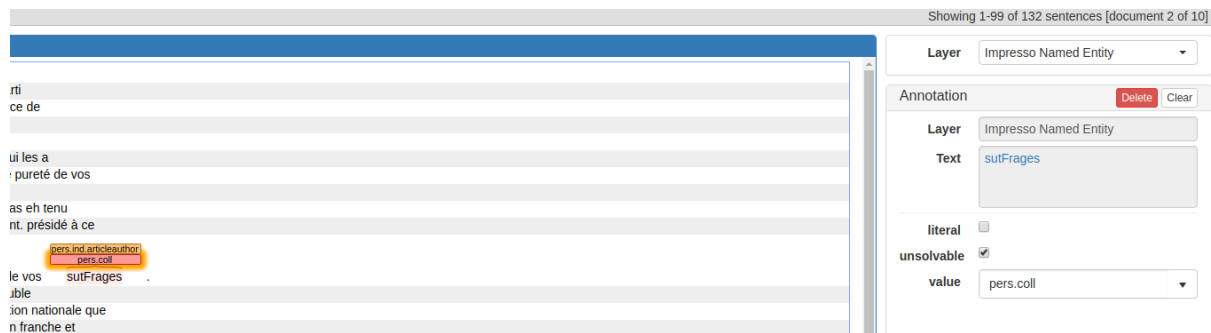
A. Unsolvable ambiguities: flag 'unsolvable'

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.



Showing 1-99 of 132 sentences [document 2 of 10]

Layer: Impresso Named Entity

Annotation: Delete Clear

Layer: Impresso Named Entity

Text: sutFrages

literal:

unsolvable:

value: pers.coll

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.

Eine Erklärung des Quai d'Orsay

Eine Erklärung des

<org>

<loc> Quai d'Orsay </loc>

</org>

Une déclaration du Quai d'Orsay

Une déclaration du

<org>

<loc> Quai d'Orsay </loc>

</org>

Die rue de Grenelle hat auf diese Aussage reagiert

Die

<org>

<loc> rue de Grenelle </loc> </org> hat

auf diese Aussage reagiert

La rue de Grenelle a réagi à cette déclaration

La

<org>

<loc> rue de Grenelle</loc>

</org> a réagi à cette déclaration

Die Élysée erklärt

Die <org><loc> Élysée </loc></org> erklärt

L'Élysée a déclaré...

L' <org><loc>Élysée </loc></org> a

déclaré...

3. Entities

3.1 Person

When the entity refers to individual or collective person (more than one individual) including fictitious persons. 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*).

A. Subtype

- `pers.articleauthor`: special 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 correspondent xx in yy*”

This is the only subtype for persons, every other person NE is annotated with `<pers>`

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
 - religious persons are not annotated in namedays and addresses

Der Bürgermeister von Paris => only 'Paris'

le maire de Paris => only 'Paris'

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

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

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

3.2 Organisations

Examples of organisations

- 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.

Die Peugeot Gesellschaft

Die `<org>`
Peugeot Gesellschaft `</org>`

Ich arbeite bei Peugeot

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

Die UNESCO

Die `<org>` UNESCO
`</org>`

Die Rote Armee

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

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

Die `<org>` Grüne Partei`</org>`

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

Die `<org>` Partei
JungsozialistInnen Schweiz
`</org>`

Die Gewerkschaft UNIA

die `<org>`
Gewerkschaft
UNIA
`</org>`

Die Gewerkschaft des Verkehrspersonals

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

La société Peugeot

La `<org>`société Peugeot`</org>`

Je travaille chez Peugeot

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

L' UNESCO

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

L'Armée Rouge

L' `<org>` Armée Rouge`</org>`

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

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

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

Le `<org>` parti socialiste
`</org>`

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

Le `<org>` parti Europe Écologie
`</org>`

Le syndicat FSU

Le `<org>` syndicat FSU `</org>`

Le syndicat national de la magistrature

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

- 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, ministry parliament, prefectures, ministries dioceses, tribunal, court, government treasury, public treasury, international org.

Die Stadtverwaltung Bern

Die <org>
Stadtverwaltung
<loc> Bern </loc>
</org>

La Mairie de Paris

La <org> mairie de
<loc> Paris </loc>
</org>

Das Bistum Basel

Das <org> Bistum
<loc> Basel </loc>
</org>

Le diocèse de Blois

Le <org> diocèse de
<loc> Blois </loc>
</org>

3.3 Locations

Examples of locations, all instinctively marked as <loc>

A. Administrative locations: refer to a territory with a geopolitical border.

- **district, city:** includes cities and all smaller units:
 - city, village, hamlet, locality, commune;
 - part of the city: district, borough, etc.

Zürich

<loc> Zürich </loc>

Der Kreis 4

Der <loc>Kreis 4 </loc>

Paris

<loc> Paris </loc>

Die Stadt Zürich

Die <loc> Stadt Zürich </loc>

La Bolline

<loc> La Bolline </loc>

Big Apple

<loc> Big Apple </loc>

Val de Crüye

<loc> Val de Crüye </loc>

Le 13e arrondissement

Le <loc> 13e arrondissement </loc>

Maison Blanche

<loc> Maison Blanche </loc>

La ville rose

La <loc> ville rose </loc>

La ville de Paris

La <loc> ville de Paris </loc>

- **region:** refers to internal divisions within a state 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> Autonome Gemeinschaft
Baskenland </loc>

la CAPS
la <loc> CAPS </loc>

Im Süden von Israel
<loc>
Im Süden von
Israel
</loc>

Au sud d'Israël
au <loc>
sud d'Israël
</loc>

Le Pays basque espagnol
Le <loc> Pays basque espagnol
</loc>

- **national:** for countries.

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

Das Vereinigte Königreich
Das <loc> Vereinigte Königreich
</loc>

Le Royaume-Uni
Le <loc> Royaume-Uni </loc>

- **supranational:** 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 Sud⁷*

Das Baskenland
Das <loc> Baskenland </loc>

Le Pays basque
Le <loc> Pays basque </loc>

Die Region um den Atlas
Die <loc>
Region um den Atlas
</loc>

La région de l'Atlas
La <loc>
Région de l'Atlas
</loc>

B. Physical places:

⁷ In the sense of the countries of the South. In other contexts, the south could designate other geographical locations (*le Sud de la France*).

- terrestrial physical locations:

Geonyms⁸ 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> Ätna </loc>

L'Étna

L' <loc> Étna </loc>

Die Wüste Gobi

Die <loc>
Wüste Gobi
</loc>

Le désert de Gobi

Le <loc>
désert de Gobi
</loc>

- aquatic physical sites:

Hydronyms⁹ refer to water bodies¹⁰, such as rivers, streams, ponds, marshes, lakes, seas, oceans, marine currents, canals, springs, etc.

Die Spree

Die <loc> Spree </loc>

La Seine

La <loc> Seine </loc>

Der Canal Saint-Martin

Der <loc>
Canal Saint-Martin
</loc>

Le Canal Saint-Martin

Le <loc>
Canal Saint-Martin
</loc>

- astronomical physical places: includes planets, stars, galaxies, etc., and their parts.

Der Mond

Der <loc> Mond </loc>

La Lune

La <loc> Lune </loc>

Die Milchstrasse

Die <loc> Milchstrasse </loc>

la mer de la tranquillité

La <loc> mer de la tranquillité
</loc>

C. Pathways:

refer to streets, squares, roads, highways, etc.

Die Autobahn A6

Die <loc>
Autobahn A6
</loc>

place de l'Abbé Georges Hénocque

<loc> place de l'

⁸ Definition taken from Mickaël Tran's thesis, Université de Tours, 2006, p. 84

⁹ Definition taken from Mickaël Tran's thesis, Université de Tours, 2006, p. 84

¹⁰ We include water streams as well.

<p><i>Die A6</i> Die <loc> A6 </loc></p> <p><i>Die Nordring Autobahn</i> Die <loc> Nordring Autobahn </loc></p> <p><i>Der Nordring</i> Der <loc> Nordring </loc></p>	<p><pers> Abbé Georges Hénocque </pers> </loc></p> <p><i>rue de Vaugirard (Vaugirard is a village)</i> <loc> rue de Vaugirard </loc></p> <p><i>la 118</i> la <loc> 118 </loc></p> <p><i>le triangle de Rocquencourt</i> le <loc> triangle de Rocquencourt </loc></p> <p><i>L'autoroute A6</i> L' <loc> autoroute A6 </loc></p> <p><i>rue des Glycines</i> <loc> rue des Glycines </loc></p>
--	---

D. Buildings :

Named buildings (train station, museum, ..) as well as their extensions (stadium, campus, university, camping...) often refer to the physical location of an organisation.

<p><i>Zürich Hauptbahnhof</i> <loc> Zürich Hauptbahnhof </loc></p> <p><i>Bern Bümpliz Nord</i> <loc> Bern Bümpliz Nord </loc></p> <p><i>Der ehemalige Bahnhof Letten</i> Der <loc> ehemalige Bahnhof Letten </loc></p> <p><i>Schloss Kyburg</i> <loc> Schloss Kyburg </loc></p> <p><i>Die Kyburg</i> Die <loc> Kyburg </loc></p>	<p><i>La gare de Rungis</i> La <loc> gare de Rungis </loc></p> <p><i>la gare Saint-Germain Grande Ceinture</i> la <loc> gare Saint-Germain Grande Ceinture </loc></p> <p><i>l'ancienne gare de Rungis</i> l' <loc> ancienne gare de Rungis </loc></p> <p><i>le palais de l'Élysée</i> Le <loc> palais de l'Élysée </loc></p> <p><i>l'Élysée</i> l' <loc> Élysée </loc></p>
--	--

E. Addresses:

- physical addresses: an address is a point in space (e.g. a point in a street)

Ich wohne in der Sihlstrasse 15 3. Stock

```
Ich wohne in der  
<loc>  
    Sihlstrasse 15 3. Stock  
</loc>
```

9 place de Rungis

```
<loc>  
    9 place de Rungis  
</loc>
```

J'habite 15 rue de Vaugirard escalier 2

```
J' habite  
<loc>  
    15 rue de Vaugirard escalier 2  
</loc>
```

31, Quai du Mont-Blanc Genova

```
<loc>  
    31, Quai du Mont-Blanc Genova  
</loc>
```

- **electronic addresses:**

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> 01 69 85 80 02 </loc>
```

mon numéro est le 01 69 85 80 02

```
mon numéro est le  
<loc> 01 69 85 80 02 </loc>
```

Mein Skype-Name ist jean.dupont

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

mon identifiant skype est jean.dupont

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

Radio Bleue auf 98.8 MHz

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

Radio Bleue sur 98.8 MHz

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

Folgt mir auf Twitter unter @leguidedannotation

```
Folgt mir auf Twitter unter  
<loc>  
    \@leguidedannotation  
</loc>
```

suivez-moi sur Twitter à @leguidedannotation

```
suivez-moi sur Twitter à  
<loc>  
    \@leguidedannotation  
</loc>
```

3.4 Human production

Media (to annotate as <prod.media>): newspapers, magazines, broadcasts, sales catalogues, etc.

(*Die Zeit; Le Figaro, Le sept à huit, La ferme célébrités*).

The name of the newspaper under annotation is annotated only when it is mentioned in the body of the newspaper articles. It is not annotated at the page-level including advertisements, images, footers, headers...

Doctrine (to ignore): political, philosophical, religious, sectarian doctrines.

(*Der Sozialismus, Theravada Buddhismus; Zeugen Jehovas; Le socialism, le bouddhisme theravâda, le structuralism, la scientology*).

Special cases for websites:

- reference to the access to the site: <loc> :
Lesen sie den Artikel auf lemonde.fr;
retrouvez cet article sur lemonde.fr
- reference to the site 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> :
Sarkozy bemängelt mediapart.fr;
Sarkozy dénonce mediapart.fr

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

3.5 Non-annotated entities

- Expressions of time (unlike in Impresso)
 - Human productions (unlike in Impresso)
 - 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*¹¹, 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>)
- we do not annotate *complex d'Oedipus* but we must annotate *Oedipus* (<pers>)
- we do not annotate *Statue of Pushkin* but we must annotate *Pushkin* (<pers>)

4. Quick guide and concrete considerations

4.1 Punctuation marks

All punctuation marks (including apostrophes) attached to named entities are left as separate tokens. They are not annotated except when they belong named entities such as for addresses, acronyms and abbreviations. Here are some examples:

(Berlin)

(<loc> Berlin </loc>)

Dr. Duboc

<per> Dr . Duboc </per>

Dr. Duboc lives in *Berlin*.

<per> Dr . Duboc </per> lives in <loc> Berlin </loc>.

H. C. Lausanne

<org>H . C . Lausanne</org>

Quai du Mont-Blanc, Geneva

<loc> Quai du Mont-Blanc, Geneva </loc>

4.2 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.

¹¹ However, this term is annotated if it refers to a group of people rather than a process, see [section 2.3.1.2](#).

B. Reporting hesitations

For any dubious cases, please report you questions with screenshot and comments at the end of this file, ideally with screenshots including context, and annotation options:

<https://docs.google.com/document/d/1yg7MGSfOvPnGoSXBWOuQaei0bY6xtTqtGTJtNuPyaeY/edit>

C. Inception mini-tutorial

<https://docs.google.com/document/d/1Tk6oadZNVVxHKSsLpVkJgOeKGS5kZTbgKxnsTpLB-vM/edit?usp=sharing>

4.3 Overview of types, subtypes and components

<i>Entity types and subtypes</i>	
<code>pers</code>	<ul style="list-style-type: none">• A single person (<i>Roger Federer</i>)• A named group of people including musical groups (<i>die Beatles, La Mano Negra</i>). (note: <i>die Schweizer, Les français</i> are not annotated.)
<code>pers.articleauthor</code>	A single person who is the author of an article.
<code>org</code>	<ul style="list-style-type: none">• Organization that markets products or provides services (<i>Die Peugeot Gesellschaft, Die Waid; La société Peugeot, la Pitié-Salpêtrière</i>). (note: <i>Die schweizer Polizei</i>; <i>la police française</i> ist not annotated)• Including special type related to newspaper to spot press agencies (a subtype for <i>Impresso v2.0</i>).
<code>loc</code>	<ul style="list-style-type: none">• District, locality, hamlet, village, city, etc. (<i>Paris, Val de Crüye</i>).• Cantons, communities of municipalities, departments, regions, etc. (<i>Autonome Gemeinschaft Baskenland; les Bouches du Rhône, Le Pays-Basque espagnol</i>).• Countries (<i>Schweiz; France</i>).• World regions, continent (<i>Maghreb; Pays-Basque</i>).• Mountains, plains, plateaus, caves, volcanoes, canyons (<i>Die Alpen, Der Vesuv; gouffre de Padirac, Le mont Ventoux</i>).• Oceans, seas, rivers, streams, ponds, marshes (<i>Der Atlantik, Der Golfstrom; La Seine, Le Lac Paladru</i>).• Planets, stars, galaxies and their parts (<i>Der Mond, Die Milchstrasse; La terre, la mer de la Tranquillité</i>)• Roads, highways, streets, avenues, squares, etc. (<i>Die Autobahn A6; L'autoroute A6</i>).• Buildings (<i>Der Prime Tower; Le Palais de l'Élysée</i>).

	<ul style="list-style-type: none"> • Physical addresses (<i>LIMSI-CNRS, Bâtiment 508, BP133, 91403 Orsay Cedex</i>). • Electronic contact information (telephone and fax numbers, URL, e-mail address, identification of social network or Internet communication tools, etc., <i>http://www.limsi.fr/, 01-69-85-80-00</i>)
prod.media	Newspapers, magazines, broadcasts, sales catalogues, etc. (<i>Die Zeit; Le Figaro, Le sept à huit, La ferme célébrités</i>).

5. Stance annotation guidelines

Stance annotation consists of deciding whether an author of a text talking about an entity in a positive/favorable or in a negative/unfavorable light, or if the statement is rather objective/neutral. Three cases of stances can thus be distinguished: two cases of subjectivity, in which case we can directly indicate the polarity (POS, NEG), and the case of non-subjectivity, objectivity or neutrality (OBJ).

OBJ is the default option, so there is no need to label neutral/objective examples.

Since stance detection is a new task, we believe that the guidelines will be enriched alongside the annotation, ambiguities will be explored gradually as tests and annotations. The more examples we have, the better it is.

In order to define a starting standard to annotate stances toward topics and named entities in a piece of text, we propose below some suggestions and clarifications that may help.

1. We are not interested in knowing author's feeling but we look for author's stance with respect to a target entity. The stance expressed towards the entity is not related to whether the whole piece of text is positive or negative.
2. We have to separate good/bad news from the stance expressed. We should NOT annotate the good/bad content of the news. E.g. if the news talks about the damage of the fire of the Notre Dame Cathedral, the stance with respect to the Cathedral is objective (OBJ), even if this is considered bad news.
3. The annotator can imagine that he is the one being talked about: would he like or dislike the statement?
4. In case of doubt, it is absolutely recommended to not mark the stance. It will be considered as OBJ, the default option.
5. If an entity X indicates the faults of another entity Y in the text, note that the stance is negative only towards Y, the stance towards X is neutral.

6. Named entity linking guidelines

Named Entity Linking (NEL) aims to disambiguate entities by linking them to entries of a Knowledge Base (KB). The following subsections provide some explanations about the annotation of named entity linking.

6.1 How Specific Should Linked Entities Be?

It is important to resolve disagreement when more than one annotation is plausible. The TAC-KBP annotation guidelines (tac, 2012) specify that different iterations of the same organization (e.g. the KB:111th U.S. Congress and the KB:112th U.S. Congress) should not be considered as distinct entities.

Example

Adams and Platt are both injured and will miss England's opening World Cup Qualifier against Moldova on Sunday. (AIDA)

Here the mention "World Cup" is labeled as KB:1998 FIFA World Cup, a specific occurrence of the event KB:FIFA World Cup. Therefore, the real entity is KB:FIFA World Cup.

6.2 Metonymy

Another situation in which more than one annotation is plausible is metonymy, which is a way of referring to an entity not by its own name but rather a name of some other entity it is associated with.

Example

Moscow's as yet undisclosed proposals on Chechnya's political future have , meanwhile, been sent back to do the rounds of various government departments. (AIDA)

The mention here, "Moscow", could be labeled as KB:Government of Russia, KB:Moscow(the city) or KB:Russia. However, neither the city nor the country can actually make a proposal. The real entity in play is KB:Government of Russia.

ANNEX A Main changes w.r.t Quaero v1

- reduction of the type of linguistic expressions considered as named entity (predominance of proper name)
- reduction of the components taken into account
- addition of 2 subtypes: pers.ind.artauthor and org.ent.pressagency
- the 2 subtypes of org.adm and org.ent are kept (w.r.t to quaero v2)

ANNEX B Main changes w.r.t Impresso v2

- New Preamble
- Removed NE types “human productions” and “time”, updated section “non-annotated entities” accordingly
- Minor additional changes in relation to the above

ANNEX C Main changes w.r.t NewsEye v3

- Removed most NE subtypes, except <pers.ind.articleauthor>, renamed <pers.articleauthor> and <prod.media>
- Briefly, our types and changes from v2 are the following:
 - <pers>: everything as in Impresso, except we ignore all subtypes (thus mark person NEs as <pers>) with one exception: <pers.articleauthor>
 - <org>: everything as in Impresso, except we ignore all subtypes (thus mark organisation NEs as <org>)
 - <loc>: everything as in Impresso, except we ignore all subtypes (thus mark organisation NEs as <loc>)
 - <loc>: everything as in Impresso, except we ignore all subtypes (thus mark organisation NEs as <loc>)
 - <prod>: we only use <prod.media>. This is our only type of <prod>.
- We removed everything related to components
- We added guidelines for the annotations of stance and for NEL