MAPA

ANONYMISATION

Annotation Guidelines for Named Entities in MAPA

Status of the document	V1.2 Revised operational version
Authors	Victoria Arranz, Chomicha Bendahman, Elena Edelman, Mickaël Rigault, Khalid Choukri (ELDA)

Contributors	Lucie Gianola, Cyril Grouin, Patrick Paroubek, Pierre Zweigenbaum (LIMSI);
	Laurent Bié, Hans Degroote, Ángela Franco (Pangeanic);
	Ona de Gibert, Maite Melero (SEAD);
	Ēriks Ajausks, Roberts Rozis, Rinalds Vīksna (Tilde);
	Albert Gatt, Mike Rosner (University of Malta);
	Montse Cuadros, Aitor Garcia (VICOMTECH)

Revised by	MAPA Consortium
Created	2020/09/15

Table of Contents

1. Introduction

This document provides the Named Entity (NE) annotation guidelines to be followed within the MAPA project for its data annotation work. For that purpose, the NE hierarchy defined within the project is described and illustrated here, with instructions on what needs to be annotated for future de-identification/anonymisation.

The objective of MAPA is to build a multilingual anonymization toolkit that can anonymize personal and sensitive data. For that purpose, multilingual language data (in all the languages to be covered by the toolkit) need to be annotated with the named entities detected, thus providing material for the development and evaluation of the system.

2. Named Entity Hierarchy

2.1 Hierarchy Levels in the NE model

The underlying model of the Named Entity hierarchy has three levels of elements. These can be seen in Figures 1 to 3. This hierarchy serves to structure both the elements to be annotated as well as those that function as semantic references or classifiers. These three levels are the following:

- **Level 1 entities** (in orange): implicit entities, they can be inferred from their annotated elements. They indicate the higher-level entities that the level-2 entities refer to and are thus used as a semantic reference to subsume the other entities.
- **Level 2 entities** (in blue): either explicit or implicit entities that may comprise some level-3 components and types to be annotated. They are also semantic classifiers for the lower level elements.
- **Entity components and types** (in green): these are either components within an entity or types of entity. They must be annotated if they have been defined within the hierarchy. Not all level-2 entities have such components.

The annotation tool INCEpTION will allow to select all entities and entity components/types to be annotated during annotation, regardless of their level within the hierarchy. This means that a word may be annotated with elements from any of the 3 levels if this is allowed by the annotation schema.

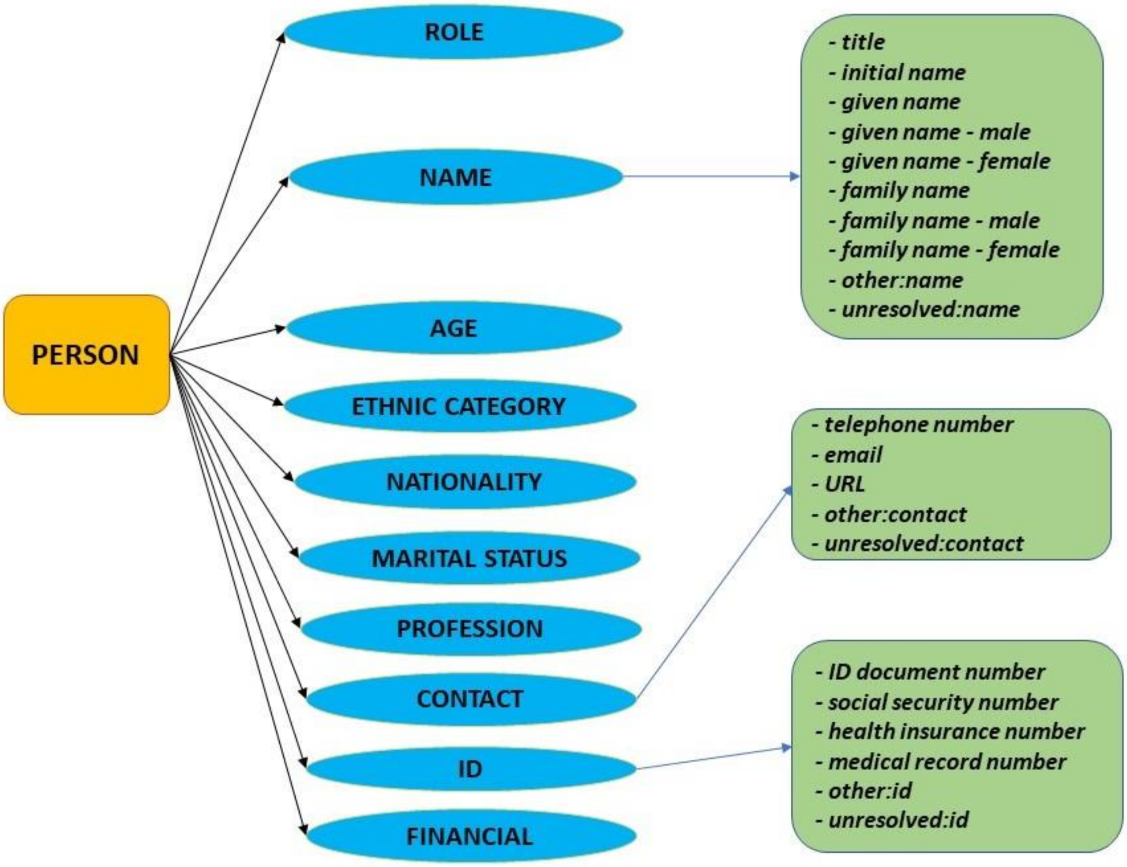


Figure 1: Named Entity Model for PERSON-related Entities

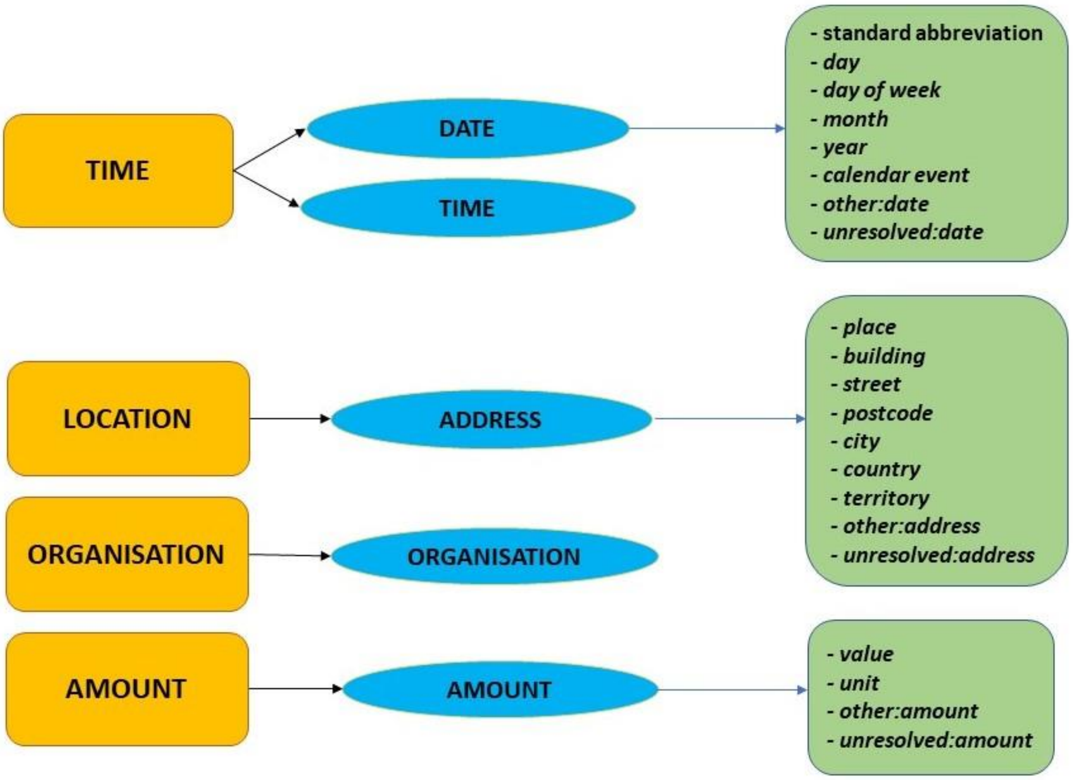


Figure 2: Named Entity Model for TIME, LOCATION, ORGANISATION and AMOUNT-related Entities

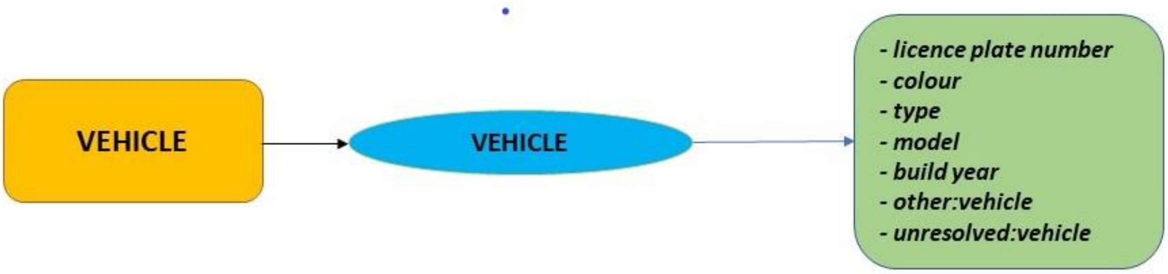


Figure 3: Named Entity Model for VEHICLE-related Entities (Legal Domain)

The NE hierarchy illustrated in Figures 1 to 2 comprises the entities addressed by the MAPA project for all domains (general, medical and legal domains). However, the VEHICLE entity in Figure 3 is only to be used in the annotation of legal-domain data.

2.2 What to Annotate within the Hierarchy

Given that several entities can be fully inferred either from their lower-level entities or from their level-3 components/types, not all elements within the hierarchy will be annotated. This would be repetitive and very time consuming. For that reason, ***the elements (entities, components and types) to be annotated are all indicated in green in Figures 4 to 6. Boxes in blue or orange will not be annotated.***

The arrows linking the different levels show two colours: the dark and thicker black arrows point to the entity relations to be annotated, while the light grey thin arrows establish the relationships that do not need to be annotated as the higher-level element is inferred from the lower-level element. This implies that:

- DATE, ADDRESS and AMOUNT will be annotated with their components/types.
- PERSON will be the only level-1 entity to be annotated directly with its level-3 components, bypassing NAME, which can be inferred from its components.
- PERSON will also be annotated when ROLE and PROFESSION take place.
- Level-2 entities NAME, CONTACT and ID will not be annotated. They will be inferred from their annotated level-3 components/types.

As indicated for the underlying model, the VEHICLE entity in Figure 6 will only be annotated in the legal domain use case.

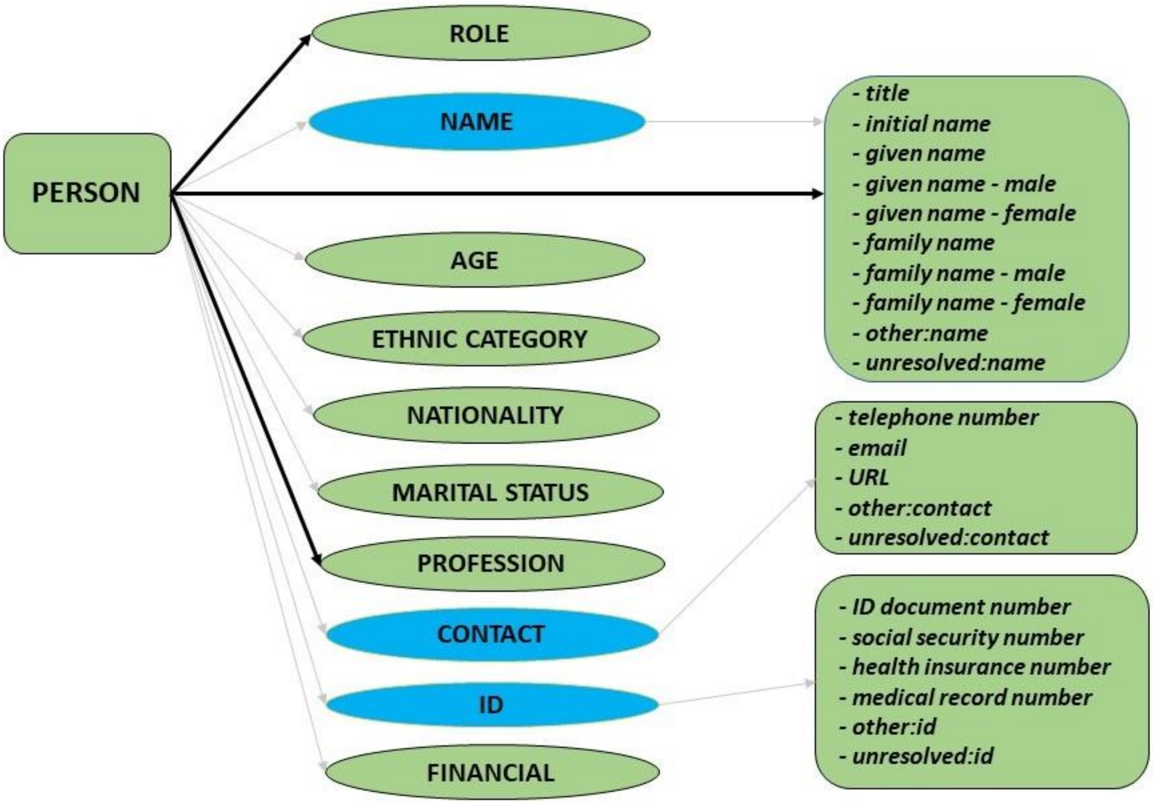


Figure 4: Named Entity Annotation Specifications: all Domains

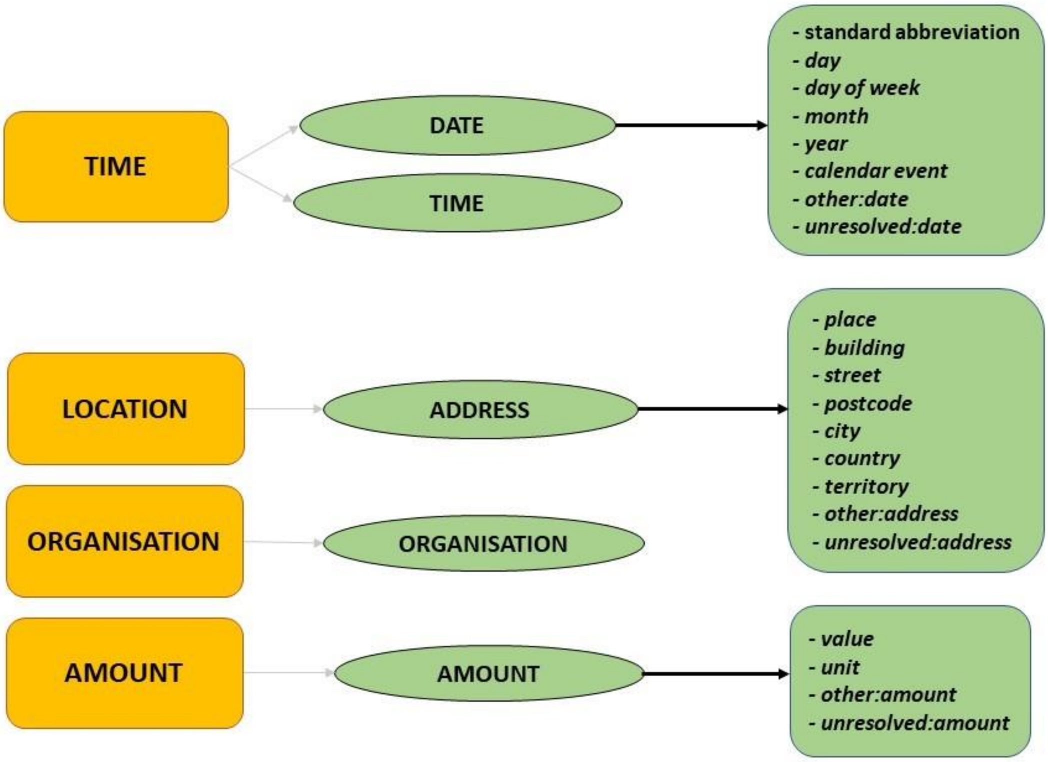


Figure 5: Named Entity Annotation Specifications: all Domains

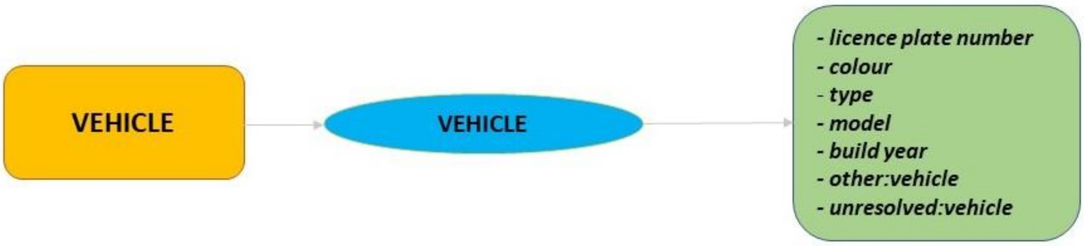


Figure 6: Named Entity Annotation Specifications: Legal-Domain Specific Entities

2.3 General Principles

This section establishes some general principles to be born in mind all throughout the annotation work:

- Annotation always needs to be done considering **the domain of the entity**, thus:
  - An element should be annotated if relevant for the domain being processed, e.g.: information on vehicles (cf. Figure 6) will be annotated only if working on the legal domain use case.
  - Some data (e.g.: EUR-LEX corpora) contain many general references like Directives, Decisions, JO documents. These are often a combination of letters and figures, with what seem to be years (dates) within the numerical references. These references and the years contained inside will not be annotated.

    Some examples follow in bold:

    - "Decision 2008/909/JHA", "Regulation No 6/2002", "Directive 97/81/EC"
    - Directive 2003/109/CE du Conseil, du 25 novembre 2003, relative au statut des ressortissants de pays tiers résidents de longue
    - Directive 2000/78/CE
    - "(OJ 1977 L 61, p. 26)", "1971 Federal Law"
    - Proposition de la Commission **[COM(1999) 566 final, p. 8]**
    - (JO 2004, L 16, p. 44)
- Annotation always needs to be done considering **the nature of the entity**, thus:
  - an element should be annotated according to its nature within the text, e.g.:
    - *Samsung* (referring to the organisation) should be annotated as **ORGANISATION**.
    - *Samsung* (referring to a smartphone) will not be annotated ("phone" is not an entity type).
    - In “Renault has opened a new factory in Brittany”: *Renault* will be annotated as **ORGANISATION**.
    - In “The thief escaped in a Renault Espace”: *Renault Espace* will be annotated as vehicle **model** (if relevant for the domain, see previous point).
    - In “I landed at Charles de Gaulle”: *Charles de Gaulle* will be annotated as a location (**place** and **ADDRESS**).
    - In “We are celebrating the 50th anniversary of Charles de Gaulle's death”: *Charles de Gaulle* will be annotated as a person's name (**given name – male** + **family name** + **PERSON**).
    - In “Dr. Alzheimer” (referring to a person): *Alzheimer* will be annotated as a family name (**family name** + **PERSON**).

        - However, in “Alzheimer's disease” (referring to the disease): *Alzheimer* should not be annotated. Disease names are not annotated.
- Sometimes we have lists of numerical references, like "Paragraphs 2,3,4", "Les considérants 27, 28, 29": such elements will not be annotated.
- Both **common and proper names** may be entities and thus annotated. As seen above, this depends on the candidate’s nature and domain, not on its grammatical category.
- Further, general entity candidates like *woman*, *man*, and collective nouns such as *commissioners, citizens*, etc. will not be annotated given that they are not associated to a particular person name.
- Annotation should be done of **self-contained elements**. Functional words, such as determiners, prepositions, or conjunctions occurring in the sentence will not be part of the annotated elements, <u>unless</u> such functional words are comprised within (are part of) the element itself. E.g.: *University of Manchester*, where preposition *of* is part of the university’s name, should be all annotated as one entity.
- **Punctuation** should not be annotated except if it belongs to the entity to be annotated. For example, *title*s like *Dr.*, *Mr.*, etc. should keep the full stop within the annotated element (if there is one, this is language dependent). This is also the case for the annotation of years in Croatian, where there is a full stop following the number, e.g.: “*2020.*”. Both number and full stop should be annotated as **year**.
- An often-ambiguous case to annotate relates to **the distinction ROLE and PROFESSION**: what may be a ROLE in some sentences may become a PROFESSION in others. This depends on the domain and the entity’s function within the sentence, whether it plays a role in the medical/legal context (ROLE) or not (PROFESSION), and whether it refers to an explicit person:
    - For instance, in a sentence like *Judge Frank MacIntosh delivered its final verdict on 7/7/2020*: *judge* would be a ROLE, while in the sentence *Franck MacIntosh, judge from the London High Court of Justice, was the witness for the accident*: *judge* would be a PROFESSION whereas *witness* would be the ROLE.
    - In cases like *Mr Smith, court clerk (**profession**), administrator (**role**):* we find both entities referring to the same person, so we should constitute a unique block with the level-1 tag PERSON.
    - In cases like the following, the selected elements will not be annotated as there is no person name with them:
        - Selection board’s decision not to include the **appellant** on the reserve list
        - The **plaintiff** was 55 years old at the time of the accident.
- Issues with the annotation of **PERSON and ORGANISATION:**
    - We may have a combination of both entities within the sentence: "**the Bruno and O'Brien cases" -> Organisation (Bruno), Person+family name (O'Brien)**
    - We may also have some ambiguous cases where organisations’ names are based on people’s names: "judgements of 13 January 2004, **Kühne & Heitz**" -> Date+day-month-year (13 January 2004), **Organisation (Kühne & Heitz)**
    - In order to help us disambiguate these cases we should look at their context within the document (and corrections will be done if we realise that we have annotated these elements wrongly so far). However, when we cannot decide

between Organisation and Person, we can either use both tags (very last option) or try to choose (even if wrong).

- Names of books, publications, concerts, fairs, tournaments, festivals, etc. are not relevant entities to be annotated. However, the dates contained within their names will be annotated and so will the locations, e.g.: *Zagrev 2020* will be annotated as: <ADDRESS><city>Zagrev</city></ADDRESS>
<DATE><year>2020</year></DATE>.
- Religious entities like Christ, God, etc. will not be annotated.
- Annotation will be done on the elements in the language of the document. Translations between parenthesis or foreign language content will not be annotated.

2.4 Annotation Definitions

- Level-2 entities will be annotated whenever they have no components or types (level-3).
- Level-2 entities with components or types are mostly annotated, except for the following:
  - NAME, CONTACT, ID and VEHICLE.
- Implicit entities (level 1) will be inferred from the annotated entities. E.g.: annotating an element as AGE or NATIONALITY (level 2) allows us to infer that we are dealing with level 1 entity PERSON. However, level-1 entity PERSON will be annotated in the following cases:
  - when NAME components (*title, given name*, etc.) are annotated, PERSON will be annotated too, so as to help us delimit the annotated entity (we may encounter several people within the same sentence);
  - this will also be the case with ROLE and PROFESSION: whenever these level-2 entities are annotated, so will PERSON, so as to associate them to the person they refer to.
    E.g.: "Jeff Eaton - **Appelant** -, v. Birthram Johnston - **Respondent**. ".
- Punctuation and words between entities:
  - If there is a comma, a hyphen or a parenthesis between the person's name and its role/profession (see example above), the punctuation mark will not be part of the annotated role or profession but it can remain under the umbrella of the PERSON block once all the entities are grouped underneath it (see Figure 7).
  - Roles and professions will not be annotated when there are other words in between them and the referred person.
- Component ***other***: this is to be used when none of the components available fits the description of the entity and we know that it should be something else.
- Component ***unresolved***: this is to be used when we do not know which entity component to assign from those which are provided (we are not able to choose).

3. Entities

This section describes the full entity hierarchy defined for MAPA. The following subsections are headed by the level-1 entities, main upper level elements, which may not be annotated but inferred from the level-2 entities that occur within the data (PERSON being the exception).

3.1 PERSON

This level-1 entity comprises the following level-2 entities:

- **ROLE**

  Definition: **ROLE** refers to the position or purpose that someone has in a situation, context, organisation.
  An element (one or several words) should be considered as ROLE when in relation to the medical or legal domains. See Sections 2.3 and 2.4 for full details. The entity ROLE will be annotated when it is next to the person it refers to. ROLE will not be annotated if there are other words between it and the referred person.

Figure 7 illustrates the annotation of ROLE in context (among other entities) with the INCEpTION tool:

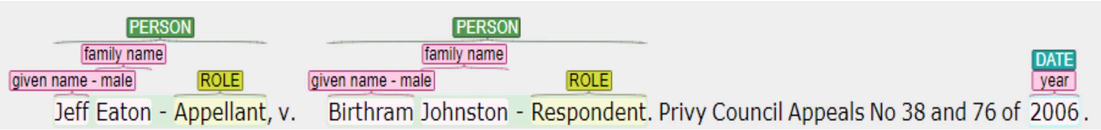

Jeff Eaton - Appellant, v. Birthram Johnston - Respondent. Privy Council Appeals No 38 and 76 of 2006.

Figure 7

- **NAME**

  Definition: This entity refers to a person's name and its title.
  Whenever possible, we will try to indicate gender within the entity's components, both for a *given name* and a *family name* (the latter are specific to some countries/cultures).
  Family names will not be assigned gender unless this is grammatically explicit within the word (such as the suffixes for non-nominative surnames in Slovak). Knowing the gender behind a well-known family name (e.g. **Castro**'s book) is not enough and that surname will not be annotated.
  Both compound *given names* and *family names* will be annotated as one single element.
  In the case of *given names*: This will also comprise middle names and patronymics. E.g.: Irina Mikhailova (first name + patronymic) will be annotated together.
  Regarding initials that a) may refer to given and family names (and which are difficult to know which tag to use) or b) may derive from some anonymization procedure, we will use the **initial name** tag (see below). All initials will use this tag, regardless of whether we know what they stand for in some cases. For example:

    - "judgements such as P. and Brock" -> Person+**initial name** (P.), Person+family name (Brock)
    - A. B. (with space): two **initial name** tags + Person tag
    - A.B. (no space): one **initial name** tag covering both non-split initials and full stops will be used + Person tag
    - A. Dupont: **initial name** + family name + Person

- AB: (no space): one **initial name** tag covering both non-split initials will be used + Person tag

Level-2 entity NAME is not annotated, but its components are. NAME's components are the following:

- **title:** this component refers to elements such as: Prof., Professor, Dr., Doctor, Mr., Mrs., military ranks, Minister, President. Titles are for life and they differ from some professions that may look like titles (e.g., rector, mayor), but that are not permanent and are associated to functions. *title* can occur as an abbreviation or as a full word (e.g.: Mr. / Mister).
  There are titles which refer to several people (e.g. MM. ou Sres.): These will be annotated as title and grouped under PERSON together with the closest person following the title. E.g.: MM Dupont et Villeroy (MM will be only associated to Dupont).
- **initial name:** this refers to the initials that may be used for both given and family names.
- **given name**: also known as *first name*. There is no gender to be indicated.
- **given name – male**: this refers to a male given name.
- **given name – female**: this refers to a female given name.
- **family name**: also known as "surname" or "last name". There is no gender to be indicated.
- **family name – male:** this refers to a male family name.
- **family name – female:** this refers to a female family name.
- **other:name**
- **unresolved:name**

Figures 7, 8, 10, 12, 13 and 14 illustrate the annotation of NAME's components in context (among other entities) with the INCEpTION tool.

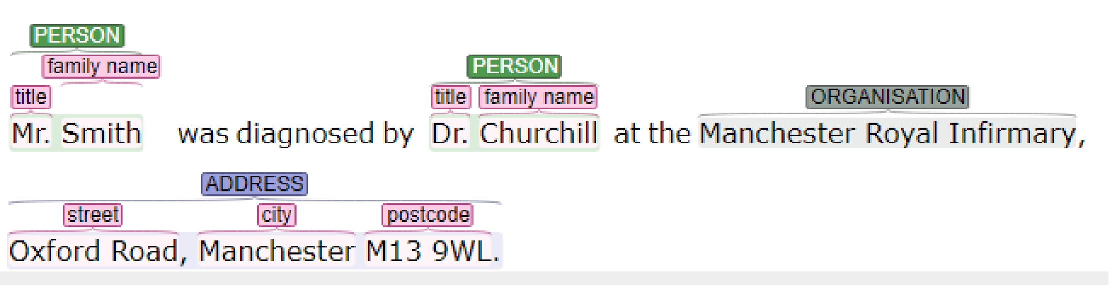

Mr. Smith was diagnosed by Dr. Churchill at the Manchester Royal Infirmary, Oxford Road, Manchester M13 9WL.

Figure 8

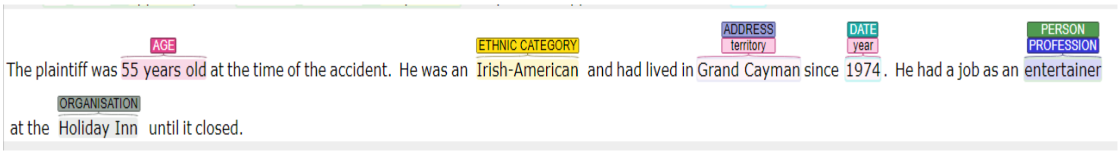
- **AGE**

Definition: This entity refers to a person's age.
When AGE is expressed in numbers (both as figures or in words), we will just annotate the number with AGE and postprocess different age ranges (see below) *a posteriori*.
Age ranges that may be postprocessed comprise the following: elderly, adult, teenager, child, newborn, underage.

Regarding age ranges like "less than 14 years old": only "14 years old" will be annotated -> less than <Age>14 years old</Age>

NB: when annotating age, the units associated should be placed within the annotated entity together with the number, e.g.: <age>55 years old</age>.

Figure 9 illustrates the annotation of AGE in context (among other entities) with the INCEpTION tool:



The plaintiff was 55 years old at the time of the accident. He was an Irish-American and had lived in Grand Cayman since 1974. He had a job as an entertainer at the Holiday Inn until it closed.

Figure 9

- **ETHNIC CATEGORY**

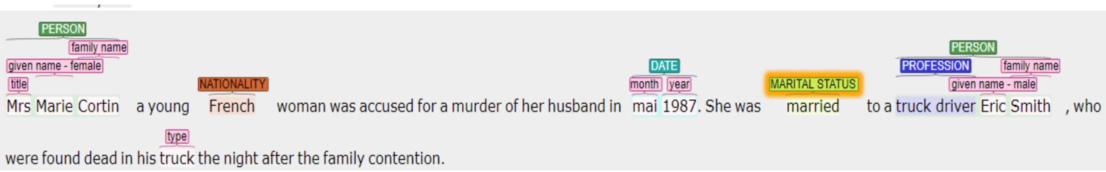
Definition: This entity refers to a number of parameters in a person's identity: race, religion, language and regional origin (sometimes known as ethno-racial, ethno-religious, ethno-linguistic, ethno-regional, respectively).
E.g.: <ethnic-category> Basque</ethnic-category>
This is an entity mostly used/found in medical and legal domains.

Figure 9 further up illustrates the annotation of ETHNIC CATEGORY in context (among other entities) with the INCEpTION tool.

- **NATIONALITY**

Definition: This entity refers to a person's demonym.
The list of nationalities considered can be consulted in Annex 1: List of Nationalities.
General NATIONALITY-like adjectives referring to geographical places will not be annotated. E.g.: **European** level, **Spanish** customs.
NATIONALITY has no components and is annotated on its own.

Figures 10 and 15 illustrate the annotation of NATIONALITY in context (among other entities) with the INCEpTION tool.



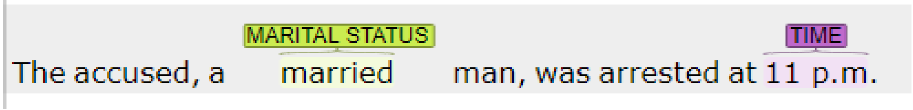
Mrs Marie Cortin a young French woman was accused for a murder of her husband in mai 1987. She was married to a truck driver Eric Smith , who were found dead in his truck the night after the family contention.

Figure 10

- **MARITAL STATUS**

  Definition: This entity is used when the following words are found: *single, married, divorced, widowed, PACS (civil solidarity pact), cohabitation*.

Figures 10 and 11 illustrate the annotation of MARITAL STATUS in context (among other entities) with the INCEpTION tool.



The accused, a married man, was arrested at 11 p.m.

Figure 11

- **PROFESSION**

  Definition: This entity considers occupations associated to functions, such as "rector" or "mayor" (as opposed to permanent *title*).

  For the distinction ROLE and PROFESSION, please refer to the explanation in Section 2.3.

  As with ROLE, PROFESSION will be used when occurring next to the person it refers to (cf. Section 2.4).

Figure 10 illustrates the annotation of PROFESSION in context (among other entities) with the INCEpTION tool.

- **CONTACT**

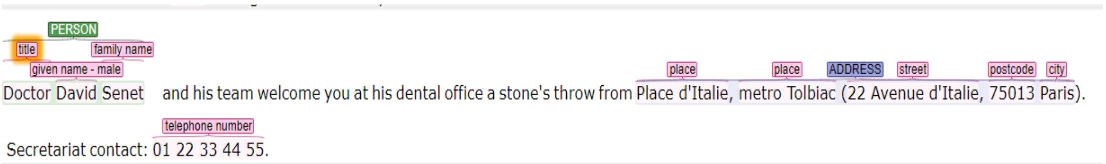
  Definition: This entity considers all contact information.

  It may have the following components/types:

  - **telephone number**
  - **email**
  - **URL**
  - **other:contact** this component may annotate information like IP address, among other types.
  - **unresolved:contact**

Only the types within CONTACT will be annotated (and not CONTACT itself). The latter can be inferred from its components/types.

Figure 12 illustrates the annotation of CONTACT information in context (among other entities) with the INCEpTION tool.



Doctor David Senet and his team welcome you at his dental office a stone's throw from Place d'Italie, metro Tolbiac (22 Avenue d'Italie, 75013 Paris).

Secretariat contact: 01 22 33 44 55.

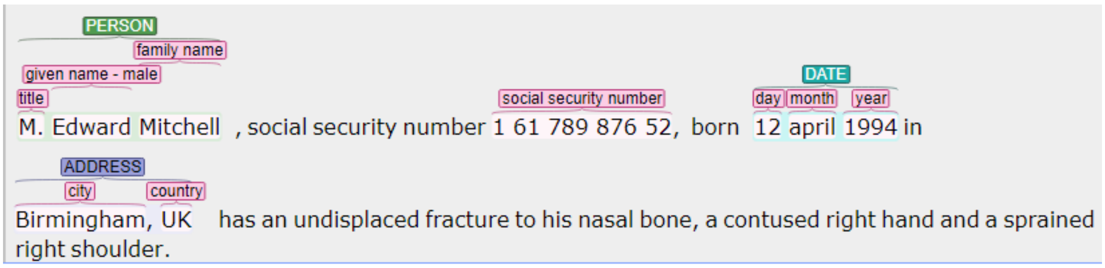
Figure 12

- **ID**

  Definition: This entity refers to different types of identification or identity numbers. It may have the following components (ID types):
  - **ID document number**: to be used for national identity, passport or driving license numbers (depending on the country).
  - **social security number**
  - **health insurance number**
  - **medical record number**
  - **other:id:** this component will cover vehicle identifier, device identifier, device serial number, legal registry numbers (where a court case is reported), etc.
  - **unresolved:id**

As with CONTACT, ID itself will not be annotated either, only its types will.

Figures 13 illustrates the annotation of ID information in context (among other entities) with the INCEpTION tool.



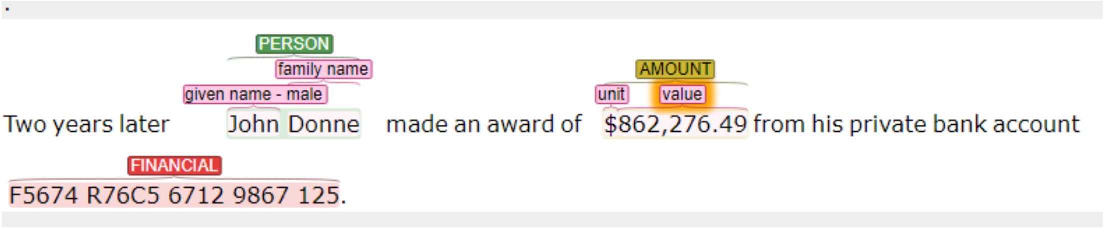
M. Edward Mitchell , social security number 1 61 789 876 52, born 12 april 1994 in Birmingham, UK has an undisplaced fracture to his nasal bone, a contused right hand and a sprained right shoulder.

Figure 13

- **FINANCIAL**

  Definition: This level-2 entity will cover all types of financial numbers: bank account, IBAN, BIC, credit card number, etc.

Figure 14 illustrates the annotation of FINANCIAL in context (among other entities) with the INCEpTION tool.



Two years later John Donne made an award of $862,276.49 from his private bank account F5674 R76C5 6712 9867 125.

Figure 14

3.2 TIME

This level-1 entity makes a distinction between entities referring to dates (DATE) and those referring to hours (TIME). It distinguishes the following level-2 entities:

- **DATE**

  Definition: DATE refers to dates, time points in the calendar, days of week, etc. Its components/types are the following:

  - **standard abbreviation**: This refers to full simplified dates (e.g.: 20/06/2020) where we will annotate the whole date block with the level-2 entity DATE without any further information (component). This also helps handle different date formats derived from different cultural origins (e.g.: US vs. UK dates).
    Otherwise, if dates are split, we will annotate their components accordingly, as listed below (e.g.: day + month + year).
  - **day:** This refers to the day of the month, it can be a numerical value (19, 20) or be represented in words (twenty).
  - **day of week:** one of the seven days in a week, including abbreviations: Monday, Tuesday, Wed, etc.
  - **month:** name of a month, including abbreviations: January, Feb, etc.
  - **year:** either a numerical value (2020) or in words (twenty-twenty)
  - **calendar event:** this refers to events such as Christmas, Easter, Halloween, etc.
  - **other: date**
  - **unresolved:date**

  Regarding dates, both level-2 DATE and its components are used for the annotation.

  Figures 7, 9, 13, 15 and 16 illustrate the annotation of DATE in context (among other entities) with the INCEpTION tool.

- **TIME**

  Definition: TIME is used for hours, in figures (e.g.: 5 a.m.) or in words (seven thirty, noon, midnight).
  This entity excludes durations (which are annotated in AMOUNT further down).
  Generic words such as **past** and **future** will not be considered as entities.

Figures 11 and 16 illustrate the annotation of TIME in context (among other entities) with the INCEpTION tool.

3.3 LOCATION

This level-1 entity comprises the following level-2 entity:

- **ADDRESS**

  Definition: ADDRESS annotates all spatial points, places, and means to locate and/or place a person. Different types of address elements can be defined. These are identified with the following components:

    - **place:** to be used for places on larger spaces/surfaces, such as a university campus, military places such as "81st infantry unit", parks (e.g.: Hyde Park), squares, etc.
    - **building:** to be used for building + building number. Here we would consider buildings like the Pentagon, as well as a specific church or religious building (e.g.: Westminster Abbey), etc., when contextualized as a location and not as an organisation.
    - **street:** to be used for street + street number.
    - **postcode:** This can cover ZIP and all other postal codes.
    - **city**
    - **country**
    - **territory:** to be used for states, counties, departments, provinces, regions, boroughs, neighbourhoods, districts, etc. There are many administrative divisions, this is country dependent.

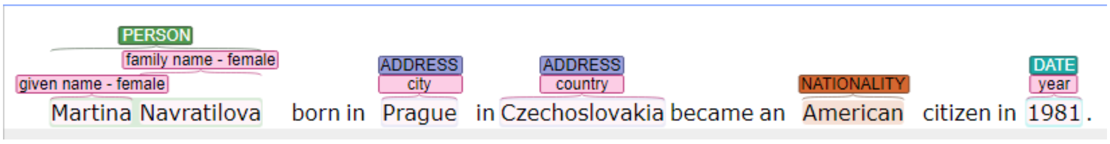
      territory will be also used for large locations like the Mediterranean Sea, the Alps, etc.
    - **other:address**
    - **unresolved: address**

    When different ADDRESS components are separated by other words that are not part of the ADDRESS and that should not be annotated, we will use several ADDRESS tags, if needed, within the sentence, to delimit ADDRESS blocks (see example in Figure 16).

    Standalone geographical locations not referring to or identifying a person should not be annotated (e.g.: There are many people in **London**).

Regarding addresses, both ADDRESS and its components will be used for the annotation.

Figures 8, 9, 12, 13, 15 and 16 illustrate the annotation of ADDRESS in context (among other entities) with the INCEpTION tool.



Martina Navratilova born in Prague in Czechoslovakia became an American citizen in 1981 .

Figure 15

3.4 ORGANISATION

This level-1 entity comprises the following level-2 entity:

- **ORGANISATION**

Definition: This entity identifies all organisations (companies, institutions, associations, etc.). These are organised groups of people with a purpose. These organisations are addressed as follows:

  - General organisations that do not refer to a particular person will not be annotated. This is the case for the following (among others): **The Minister for Justice and Equality, The Commissioner An Garda Síochána (Irish police), Verwaltungsgerichshof (Cour administrative, Autriche), European Commission, European Parliament, Council, The Austrian and German governments, High Court, WRC, The Catholic Church.**
  - However, we do annotate such organisations if directly associated to a person's name within the sentence: *European Council officer John Doe has been arrested.*
  - Even if general organisations are not annotated, we do annotate address information (city and country, if available) if comprised within such expressions (e.g.: Cour administrative, **Autriche**). Needdless to say that this will be language dependent.
  - We will annotate other specific Organisations such as companies and places where people work, law firms, names of companies involved in court cases, etc.

Figures 8 and 9 illustrate the annotation of ORGANISATION in context (among other entities) with the INCEpTION tool.

3.5 AMOUNT

This level-1 entity comprises the following level-2 entity:

- **AMOUNT**

  Definition: AMOUNT is used to cover distances (e.g.: 3 km), quantities (e.g.: 24 kg; 126 Euros) and durations (e.g.: 12 hours, 5 and half hours) that are relevant to the domain and context of the document (medical or legal).

  Therefore, general amounts such as **200 firemen**, **30 workers**, or durations like “The commission will come up with a proposal within **6 months** after the date of the submission” will not be annotated.

  The components/types for AMOUNT are the following:

  - **value:** to be used for the numerical amount denoted (in numbers or in words). E.g.: *24* is the value in “24 kg”). Generic words such as “each” and relative ones such as “several” will not be considered as values and thus will not be annotated.
  - **unit:** to be used for the kind of amount the value refers to. E.g.: *kg* is the unit in “24 kg”.
  - **other:amount**
  - **unresolved:amount**

Values can occur as numbers or words, and units can be found in full words (kilometers) or in abbreviations (km).
Annotation hint: when units and values occur with no space between them (e.g.: $136), INCEpTION allows to annotate per character (see INCEpTION's settings and user guide), thus assigning both **unit** and **value** tags to one single block.

Figure 14 illustrates the annotation of AMOUNT in context (among other entities) with the INCEpTION tool.

3.6 VEHICLE

This level-1 entity will only be treated when dealing with data from the legal domain (it is a use-case specific entity). It comprises the following level-2 entity:

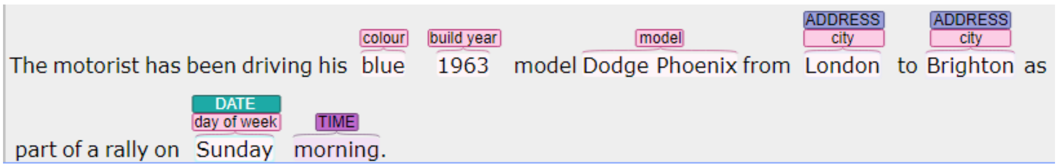
- **VEHICLE**

  Definition: VEHICLE refers to an object used for transporting people or goods, such as a car, lorry, motorcycle, bus, etc.
  Its components are the following:

  - **licence plate number**: a vehicle's registration number as indicated in its licence plate / number plate.
  - **colour:** this refers to the vehicle's colour.
  - **type:** this refers to the type of vehicle (e.g.: car, lorry, motorcycle, etc.).
  - **model:** this refers to the vehicle's model information, e.g.: Renault Clio.
  - **build year:** this refers to the vehicle's construction year, often provided with the model.
  - **other:vehicle:** this component may comprise identification numbers such as the VIN number, etc.
  - **unresolved:vehicle**

  Level-1 and level-2 entities VEHICLE are not annotated. These are inferred from their components, which are annotated.

  Figure 16 illustrates the annotation of VEHICLE information in context (among other entities) with the INCEpTION tool.



The motorist has been driving his blue 1963 model Dodge Phoenix from London to Brighton as part of a rally on Sunday morning.

Figure 16

Annex 1: List of Nationalities

Below follows a list of nationalities and their respective countries, as listed by the World Atlas (https://www.worldatlas.com/articles/what-is-a-demonym-a-list-of-nationalities.html):

Country	Demonym
Afghanistan	Afghan
Albania	Albanian
Algeria	Algerian
Andorra	Andorran
Angola	Angolan
Antigua and Barbuda	Antiguan or Barbudan
Argentina	Argentine
Armenia	Armenian
Australia	Australian
Austria	Austrian
Azerbaijan	Azerbaijani, Azeri
The Bahamas	Bahamian
Bahrain	Bahraini
Bangladesh	Bengali
Barbados	Barbadian
Belarus	Belarusian
Belgium	Belgian
Belize	Belizean
Benin	Beninese, Beninois
Bhutan	Bhutanese
Bolivia	Bolivian
Bosnia and Herzegovina	Bosnian or Herzegovinian
Botswana	Motswana, Botswanan
Brazil	Brazilian
Brunei	Bruneian
Bulgaria	Bulgarian
Burkina Faso	Burkinabé
Burma	Burmese
Burundi	Burundian
Cabo Verde	Cabo Verdean
Cambodia	Cambodian
Cameroon	Cameroonian
Canada	Canadian
Central African Republic	Central African
Chad	Chadian
Chile	Chilean
China, People's Republic of	Chinese
Colombia	Colombian
Comoros	Comoran, Comorian
Congo, Democratic Republic of the	Congolese
Congo, Republic of the	Congolese
Costa Rica	Costa Rican

Côte d'Ivoire	Ivorian
Croatia	Croatian
Cuba	Cuban
Cyprus	Cypriot
Czech Republic	Czech
Denmark	Danish
Djibouti	Djiboutian
Dominica	Dominican
Dominican Republic	Dominican
East Timor	Timorese
Ecuador	Ecuadorian
Egypt	Egyptian
El Salvador	Salvadoran
Equatorial Guinea	Equatorial Guinean, Equatoguinean
Eritrea	Eritrean
Estonia	Estonian
Ethiopia	Ethiopian
Fiji	Fijian
Finland	Finnish
France	French
Gabon	Gabonese
Gambia, The	Gambian
Georgia	Georgian
Germany	German
Ghana	Ghanaian
Gibraltar	Gibraltar
Greece	Greek, Hellenic
Grenada	Grenadian
Guatemala	Guatemalan
Guinea	Guinean
Guinea-Bissau	Bissau-Guinean
Guyana	Guyanese
Haiti	Haitian
Honduras	Honduran
Hungary	Hungarian, Magyar
Iceland	Icelandic
India	Indian
Indonesia	Indonesian
Iran	Iranian, Persian
Iraq	Iraqi
Ireland	Irish
Israel	Israeli
Italy	Italian
Ivory Coast	Ivorian
Jamaica	Jamaican
Japan	Japanese
Jordan	Jordanian

Kazakhstan	Kazakhstani, Kazakh
Kenya	Kenyan
Kiribati	I-Kiribati
North Korea	North Korean
South Korea	South Korean
Kuwait	Kuwaiti
Kyrgyzstan	Kyrgyzstani, Kyrgyz, Kirgiz, Kirghiz
Laos	Lao, Laotian
Latvia	Latvian, Lettish
Lebanon	Lebanese
Lesotho	Basotho
Liberia	Liberian
Libya	Libyan
Liechtenstein	Liechtensteiner
Lithuania	Lithuanian
Luxembourg	Luxembourg, Luxembourgish
Macedonia, Republic of	Macedonian
Madagascar	Malagasy
Malawi	Malawian
Malaysia	Malaysian
Maldives	Maldivian
Mali	Malian, Malinese
Malta	Maltese
Marshall Islands	Marshallese
Martinique	Martiniquais, Martinican
Mauritania	Mauritanian
Mauritius	Mauritian
Mexico	Mexican
Micronesia, Federated States of	Micronesian
Moldova	Moldovan
Monaco	Monégasque, Monacan
Mongolia	Mongolian
Montenegro	Montenegrin
Morocco	Moroccan
Mozambique	Mozambican
Namibia	Namibian
Nauru	Nauruan
Nepal	Nepali, Nepalese
Netherlands	Dutch, Netherlandic
New Zealand	New Zealand, NZ, Zelanian
Nicaragua	Nicaraguan
Niger	Nigerien
Nigeria	Nigerian
Northern Mariana Islands	Northern Marianan
Norway	Norwegian
Oman	Omani
Pakistan	Pakistani

Palau	Palauan
Palestine	Palestinian
Panama	Panamanian
Papua New Guinea	Papua New Guinean, Papuan
Paraguay	Paraguayan
Peru	Peruvian
Philippines	Filipino, Philippine
Poland	Polish
Portugal	Portuguese
Puerto Rico	Puerto Rican
Qatar	Qatari
Romania	Romanian
Russia	Russian
Rwanda	Rwandan
Saint Kitts and Nevis	Kittitian or Nevisian
Saint Lucia	Saint Lucian
Saint Vincent and the Grenadines	Saint Vincentian, Vincentian
Samoa	Samoan
San Marino	Sammarinese
São Tomé and Príncipe	São Toméan
Saudi Arabia	Saudi, Saudi Arabian
Senegal	Senegalese
Serbia	Serbian
Seychelles	Seychellois
Sierra Leone	Sierra Leonean
Singapore	Singapore, Singaporean
Slovakia	Slovak
Slovenia	Slovenian, Slovene
Solomon Islands	Solomon Island
Somalia	Somali
South Africa	South African
South Sudan	South Sudanese
Spain	Spanish
Sri Lanka	Sri Lankan
Sudan	Sudanese
Suriname	Surinamese
Swaziland	Swazi
Sweden	Swedish
Switzerland	Swiss
Syria	Syrian
Tajikistan	Tajikistani
Tanzania	Tanzanian
Thailand	Thai
Timor-Leste	Timorese
Togo	Togolese
Tokelau	Tokelauan
Tonga	Tongan
Trinidad and Tobago	Trinidadian or Tobagonian

Tunisia	Tunisian
Turkey	Turkish
Turkmenistan	Turkmen
Tuvalu	Tuvaluan
Uganda	Ugandan
Ukraine	Ukrainian
United Arab Emirates	Emirati, Emirian, Emiri
United Kingdom of Great Britain and Northern Ireland	UK, British
United States of America	United States, U.S., American
Uruguay	Uruguayan
Uzbekistan	Uzbekistani, Uzbek
Vanuatu	Ni-Vanuatu, Vanuatuan
Vatican City State	Vatican
Venezuela	Venezuelan
Vietnam	Vietnamese
Yemen	Yemeni
Zambia	Zambian
Zimbabwe	Zimbabwean