Named Entity Annotation Guidelines

These guidelines have been updated and modified starting from those presented at DHBenelux 2024 (https://doi.org/10.5281/zenodo.11366870).

Authors: Chiara Palladino, Margherita Fantoli, Evelien de Graaf, Matteo Romanello, Marijke Beersmans, Laura Soffiantini, Eleonora Litta, Valeria Boano, Pietro Zaccaria, Rachel Milio

Summary

Purpose of these Guidelines

The two main purposes of these Guidelines are:

1. To design a tagset and annotation strategy that could support the creation of consistent annotated datasets of Named Entities in ancient texts, minimizing annotator disagreement and ambiguity. The primary application is (not limited to) the development of Machine Learning methods.
2. To ensure a basic level of interoperability and exchange across projects that involve Named Entities in Ancient Greek and Latin, by providing labels that are both specific to ancient documents and mappable onto commonly used concepts in NLP for modern datasets.

Key Definitions

Named Entity

Named Entities in these guidelines are defined as follows:

1. A Named Entity is a string of text that acts as a designator for a referent, denoting a unique and identifiable object in the world. The main designator is given priority in labeling: if it is a personal name, it is labelled as **person**, even if other qualifiers appear together (e.g. “Thucydides the Athenian” is labelled only as **person**, since the personal name is the main designator).
2. A Named Entity is identified as the shortest possible string that designates a referent. Common nouns and qualifiers, without which the string would still function as a designator, should not be annotated (e.g. “Thucydides the Athenian”, not “the great writer Thucydides the Athenian”).
3. A named entity contains, by definition, at least one capitalized word (with very few exceptions).
4. When irrelevant to the meaning of the Named Entity,[1] articles are not annotated.

These definitions are meant to provide a pragmatic approach to annotation, in addition to some conceptual guidance to work with edge cases. However, they are conventional and will inevitably produce some outliers: wherever an indication is given that may seem in contradiction with these definitions, an explanation is provided to justify it.

Labels and Tags

This tagset provides a set of semantic labels to classify Ancient Named Entities. The tags provided include two hierarchical levels.

[1] Exceptions happen where the annotation of articles cannot be avoided, particularly in long strings such as ‘Περὶ λέξεως καὶ τῶν λεγομένων’ (**work**), ‘Ἀλέξανδρον τὸν Πριάμου’ (**person**), or in ancestry expressions, such as ‘ὁ μέν Πριάμου’ (the son of Priam - **person.ancestry**).

- **First-level tags** represent the type of object being designated by the named entity. The usage of first-level tags is intended to be strict, to ensure interoperability.
- **Second-level tags** are designed with more consideration for the semantic function of the name: they may point to the way in which a referent is designated (for example, an individual identified with an epithet rather than their own name) or the specifically identifiable subgroup to which an entity belongs (e.g. a political or racial collective). They are designed to be used, expanded, or selected for project-specific goals and contexts.

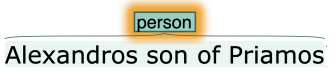
Key Rules

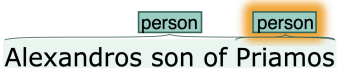
Labels and Nesting

Only the surface level of an entity is annotated. Avoid metonymic or metaphoric readings. Examples: 'I am going to Julia' (i.e. to Julia's house): 'Julia' should be annotated as a **person**, not a location; 'Athens voted for the war': 'Athens' is annotated as a **place**, even though it is intended as a collective.

Multiple second-level tags are not permitted. Only one second-level tag should be selected.

Nested annotations happen when a Named Entity is annotated inside another Named Entity. Nesting is not recommended by these guidelines. At the end of this document we will provide some recommendations on nesting in compliance with this tagset, for project-specific purposes.

The correct annotation form according to these guidelines: 

An example of an (incorrect) nested annotation: 

Annotation Boundaries

Sometimes it can be tricky to establish the boundaries of what exactly constitutes a Named Entity. These operative questions are suggested:

1) What referent (= uniquely identifiable thing) does the string talk about?
2) Can the referent be named or otherwise associated with a unique thing in the world, or in an index, without additional context?
3) What is the shortest version of the string that can be associated with the referent, without additional context?

Consider the following examples:

1) ‘Thucydides the Athenian’: the string contains two capitalized words (the person Thucydides and the ethnic Athenian) but has only one referent, Thucydides. Therefore, it is annotated fully and tagged as **person**.
2) ‘Augusti soror’ (the sister of Augustus). Even though Augustus is the name, the full string is associated with a different nameable referent, Octavia the Younger. So it should be annotated fully and tagged as **person.ancestry** (i.e., person identified by means of their ancestry). Similarly, Greek examples of patronymic expressions like ‘ὁ Πριάμου’ (the son of Priam) should be annotated in full as **person.ancestry**, including the article.
3) ‘Junonis templum’ (temple of Juno), ‘Ἀθηναίων πόλις’ (city of the Athenians). Even though ‘Juno’ is a person and ‘Ἀθηναίων’ is a collective, the strings point to an identifiable entity that can be associated with an index entry or place-name. So they should be annotated fully and tagged as **place.**
4) ‘Black sea’ and ‘upper Egypt’ are full strings referring to a **place**. ‘Black’ or ‘Egypt’ are not sufficient to identify the referent being talked about, so lowercase words must be included.

Conversely, consider the following cases:

1) ‘Mediterranean sea’: the word ‘sea’ does not further identify the entity ‘Mediterranean’, nor does it change its label. Therefore, only ‘Mediterranean’ should be tagged as **place**.
2) ‘an Athenian woman’: this expression does not point to a precisely identifiable referent, as the noun ‘woman’ is not a rigid designator for an individual. In these cases, only the ethnonym ‘Athenian’ is annotated as **person.ethnic**, as it is the only designator in the expression. See below the instructions on how to use the second-level tag **ethnic**.
3) ‘Ἰβηρικὴ παραλία’ (Iberian coast): in this expression, ‘coast’ does not point to a place that can be named or found in an index. Only the derivative adjective Ἰβηρικὴ points to an identifiable place (Iberia): therefore, only the adjective is annotated, using **place.derivative**. See below the instructions on how to use the second-level tag **derivative**.

Exceptions and Edge cases

Non-consecutive entities are strings that contain one named entity but are split across the text. These are provisionally tagged as **non-consecutive**.

Exceptional lowercase names: There are some edge cases of strings that unequivocally refer to specific entities with referents, but that appear irregularly capitalized, e.g. capitalized in certain editions but not others, capitalized only in the translation but not in the original, etc. In these cases, provided that there is secure identification of the referent, these strings should be annotated and tagged appropriately.

Examples: ‘symplegades’ (i.e. the Symplegades), ‘erynes’ (the Furies), ‘the hundred-handed’ (the Hundred-Handed), names of winds/compass points (‘septentrio’, ‘caurus’, ‘austrus’).

First-level tags

Person

Any identifiable single individual, including deities and anthropomorphic mythological figures.

Note: If the person's name appears with an epithet, patronymic, ethnic, or other attributive indication, they should be annotated together and tagged as **person**, as the proper name is the main designator in the string.

Examples: 'Zeus', 'Ἄρτεμις Λιμνάτις', 'L. C. Sulla', 'Ajax Telamonius', 'Agrippina Maior', 'Apollodorus Kepotyrannos', 'Thucydides of Athens', 'Thucydides the Athenian', 'Heraclides Ponticus', 'Ἡρόδοτος Ἁλικαρνησσεύς', 'Alexander son of Priamus', 'Ἰοῦς τῆς Ἀργείης', 'Claudia Marci filia'.

Available second-level tags: .author, .ancestry ('Ἀτρείδης'), .epithet, .ethnic, .derivative.

Place

A politically, culturally, or geographically defined location, including fictional spaces and structures like temples, buildings, specific urban areas (e.g., gymnasia), and houses.

Note: When the name appears together with various lowercase words, these should be included when they help disambiguate the entity (e.g. 'Red sea' but 'Mediterranean sea'). See Annotation Boundaries.

Examples: 'Asia', 'Περσίς', 'Athenae', 'Carthago Nova', 'city of the Athenians', 'Black sea', 'upper Egypt', 'temple of Apollo', 'delubrum Iovis Optimi Maximi', 'Τάρταρος', 'Cynosarges'.

Available second-level tags: .astronomy, .epithet ('(the) Rugged' (=Ithaka)), .derivative.

Collective

A named group of people or other creatures with shared identifiable characteristics on social, intellectual, political, national, family, mythical, or ethnic basis.

Examples: 'Ἀθηναῖοι', 'Egyptians', 'Cyclopes', 'Eumolpidae', 'Peripatetics', 'Tarquinii'.

Available second-level tags: .ancestry ('Δαναοί', 'sons of Priamus'), .animal ('cattle of Helios'), .astronomy ('Pleiades'), .ethnic ('Ἀθηναῖοι', 'Romana gens'), .organization ('Στωικοί', 'Senatus'), .epithet ('Eumenides'), .derivative ('Pygmeian', 'Academic').

Creature

Mythical or real precisely identifiable non-human, non-anthropomorphic creatures.

Examples: ‘Chiron’, ‘Polyphemus’, ‘Βουκέφαλος’, ‘Chimaera’, etc.: **.creature**.

Available second-level tags: .animal (‘Ἄργος’), .astronomy.

Event

Significant named events identified by a string with a precise boundary.

Examples: ‘Flood of Deucalion’, ‘Battle of Naupactus’, ‘*Catilinae coniuratio*’, ‘πόλεμος ὁ Ἀθηναίων καὶ Λακεδαιμονίων’, ‘*bellum Mithridaticum*’.

Language

Languages and dialects clearly identified as such.

Examples: ‘Latin’, ‘Greek’, ‘Phoenician’, ‘Oscan’, ‘Ἑλληνισμός’.

Object

Artifacts or groups of artifacts clearly identified with a name, such as ships, weapons, statues, columns, dedications, etc.

Examples: ‘Argos’ (the ship of the Argonauts, not the person, the city, or Odysseus’ dog), ‘Athena’ (the statue of Pheidias, not the goddess).

Miscellaneous

Entities that do not (yet) have a specific first-level tag among those provided.

Examples: ‘Septentrio’, ‘Caurus’, ‘Austrus’, etc.: **miscellaneous**.

Time

Any absolute date or time expression.

Examples: ‘23rd Olympiad’, ‘the second year of the 23rd Olympiad’, ‘Consulate of Cicero’, ‘Archonship of Cleisthenes’, ‘Νικοστράτου ἐπιστάτου’, ‘anno septimo tribuniciae potestatis Vespasiani’.

Work

Titles of literary or non-literary works, in any form.

Note: These guidelines do not address complex and edge cases, as these have been widely standardized by other scholars: for guidance on more project-specific needs, see Romanello & Najem-Meyer (2022) and Berti (2024).

Examples: 'Odyssey', '*Aeneis*', 'Πολιτεία', '*De Rerum Natura*', '*Oedipus Rex*', '*Noctes Atticae*', '*Anabasis of Alexander*', 'Περὶ λέξεως', '*History of the Peloponnesian War*', '*On Poets*', '*On Islands*'.

Second-level tags

Second-level tags provide further specifications for particular types of Named Entities.

.ancestry (collective.ancestry, person.ancestry)

A designation or expression that refers unambiguously to one individual or group of individuals by using a family name, patronymic, matronymic, or other indication of lineage or familial relationships.

Note: This tag should only be used when the ancestry appears in isolation. If the ancestry appears together with a name, e.g. 'Achilles Pelides', they should be annotated together with the appropriate first-level tag (e.g., **person**).

Examples: 'Tarquinii', 'Μερμνάδας', 'sons of Priamos': **collective.ancestry**; 'Atreid', 'Hector's wife', 'παῖς Ἀλκαίου', 'Saturnia' (Juno), 'ὁ Πριάμου' (Paris): **person.ancestry**.

.animal (collective.animal, creature.animal)

A type of creature or collective of creatures clearly identifiable with an animal or animal species.

Examples: 'cattle of Helios': **collective.animal**, 'Βουκέφαλος': **creature.animal**.

.astronomy (creature.astronomy, collective.astronomy, place.astronomy)

Named stars, groups of stars, constellations, and planets.

Examples: 'Pleiades': **collective.astronomy**.

.author (person.author)

A person clearly mentioned in relation to works they have authored. This tag may be modified or even omitted for project-specific goals.

Examples: 'Euripides of Athens': **person.author**.

.derivative (collective.derivative, person.derivative, place.derivative)

An adjective derived from a toponym, personal name, or group name, used to identify things that are **not** individuals or collectives (for individuals or collectives, see **.ethnic**). Only the derivative is annotated, as the common noun in the expression does not act as a rigid designator. The first-level tag depends on the name from which the adjective derives (e.g. "Iberian" will be a **place**, "Platonic" a **person**, etc.).

Note: This tag is used when a common noun in an expression is not a rigid designator, i.e. does not point to an identifiable entity ('spear', 'coast', 'book', 'school', etc.), and the only element of identification is the membership of some geographical, political, personal or otherwise defined entity (see Annotation Boundaries).

Examples: 'Iberian (coast)', 'Aegean (wind)', 'Athenian (Even the Sunthat rules the world was captive made of Love.ship)', 'Αἰγυπτιά (φόρτια)', 'Ἑλληνικῶς', 'Ὀλύμπια (δώματα)', 'Tyrias (arces)', 'Tiberina (ostia)': **place.derivative**, 'Platonic (concept)': **person.derivative**, 'Pygmeian (spears)', (**collective.derivative**).

.ethnic (collective.ethnic, person.ethnic)

An ethnonym, demonym, or other word used to identify persons or collectives by means of their membership to a geographically or ethnically defined group. This tag is *exclusively* used with persons or collectives, as ethnics are mainly used in the ancient world to identify individuals via ethnic memberships (see also our rationale below). For all other uses of adjectives derived from places, use the **.derivative** subtag.

Note: This tag should only be used when the ethnic appears in isolation, functioning as the main designator of an entity. When it appears together with the personal name, the expression is only annotated with the appropriate first-level tag, because it is the name that acts as the main designator.

- 'Athenian woman' > **person.ethnic.** The noun 'woman' does not point to a unique individual that can be identified, so only the ethnic Athenian is annotated.
- 'Thucydides the Athenian' > **person**. The personal name functions as the main designator.

Examples: 'Romana gens' (**collective.ethnic**), 'Ἀθηναῖος', 'Λακεδαιμονίης (γυναικός)' (**person.ethnic**), 'Φοίνικας', 'Δωριέας' (**collective.ethnic**).

.epithet (collective.epithet, person.epithet, place.epithet)

A *capitalized* epithet used to refer unambiguously to one individual, location, or collective, including nicknames, titles, and other appellatives.

Note 1: This tag should only be used when the epithet appears in isolation, functioning as the main designator of an entity. When it appears together with the personal name, the expression is only annotated with the appropriate first-level tag, because it is the name that acts as the main designator.

- 'Apollo Archegetes' > **person**.
- 'Soter' > **person.epithet** (the referent is Ptolemy I).

Note 2: Non-capitalized epithets with attributive function (e.g. 'swift-footed') are not addressed here. Edge cases that should be checked against dictionaries or authority lists.

Examples: 'Archegetes' (Apollo), 'Soter' (Ptolemy I), 'Γλαυκῶπις' (Athena): **person.epithet**.

.organization (collective.organization)

Collectives identified by precise organizational structures, such as priesthoods, legions, religious, intellectual, or political groups and institutions, and so on.

Examples: ‘Senatus’, ‘Βουλή’, ‘Academics’, ‘Herophileans’: **collective.organization**.

Addendum on subtags “ethnic” and “derivative”

The subtags “ethnic” and “derivative” constitute an exception to the notion of *referent* as a uniquely identifiable object: “Athenian woman” does not have an identifiable referent that can be named. However, especially in the Greek usage (but also in most ancient societies), distinctions of ethnic/racial nature are standard to identify individuals via their membership to a non-familial group with defined characteristics. In the ancient Greek world, ethnics used for persons tend to be grammatically different forms from derivative adjectives (the so-called ktetics). “Ktetics”, or “possessive” forms, characterize objects, events, languages, places and other non-human things via association. So, while an ethnic proper is part of the system of identification of a *person*, all other derivative forms do not have the same function and simply serve as adjectives: in the string “Iberian coast”, the place *coast* is not identified through association with a racial group, but as part of another location through a derivative. Based on this distinction, we are restricting the usage of ethnics to human individuals and groups.

References: P. M. Fraser, Greek Ethnic Terminology, Oxford, New York 2010; M. H. Hansen, City-Ethnics as Evidence for Polis Identity, in M. H. Hansen (ed.), More Studies in the Ancient Greek Polis, Stuttgart, 1996, 169–196; M. H. Hansen, The Use of Subethnics as Part of the Name of a Greek Citizen of the Classical Period: The Full Name of a Greek Citizen, Once Again. Studies in the Ancient Greek Polis (2004), 117–130; W. Dittenberger, Ethnika und verwandtes. I, Hermes 41 (1906), 78–102; E. Risch, Zur Geschichte der griechischen Ethnika, Museum Helveticum 14 (1957), 63–74.

References

Authority lists useful for consultation

Theoi: https://www.theoi.com/
WikiData: https://www.wikidata.org/wiki/Wikidata:Main_Page
Trismegistos: https://www.trismegistos.org/index.php
Pleiades: https://pleiades.stoa.org/
Manto Project: https://manto.unh.edu/
ToposText: https://topostext.org/index.php
World Historical Gazetteer: https://whgazetteer.org/search/#
Platner, *Topographical dictionary of ancient Rome* on Perseus
Smith, *Dictionary of Greek and Roman Geography* (1854) on Perseus

Smith, *Dictionary of Greek and Roman biography and mythology* on Perseus
Smith, *Dictionary of Greek and Roman antiquities* (1890) on Perseus
Lexicon of Greek Personal Names: https://www.lgpn.ox.ac.uk/
Romans 1by1: http://romans1by1.com/rpeople/people
Mapping Ancient Polytheisms: https://map-polytheisms.huma-num.fr/?lang=en

Bibliography

Álvarez-Mellado, Elena, María Luisa Díez-Platas, Pablo Ruiz-Fabo, Helena Bermúdez, Salvador Ros, and Elena González-Blanco. "TEI-Friendly Annotation Scheme for Medieval Named Entities: A Case on a Spanish Medieval Corpus." *Language Resources and Evaluation* 55, no. 2 (June 2021): 525–49. https://doi.org/10.1007/s10579-020-09516-2.

Berti, Monica. "Digital Catalogs of Ancient Greek Authors and Works through Papyrological Data." In *Digital Papyrology III: The Digital Critical Edition of Greek Papyri: Issues, Projects, and Perspectives*, edited by Nicola Reggiani, 89–106. De Gruyter, 2024. https://doi.org/10.1515/9783111070162-006.

Berti, Monica. "Digital Canons and Catalogs of Fragmentary Literature." In *Fragmente einer fragmentierten Welt*, 217–36. De Gruyter, 2024. https://www.degruyterbrill.com/document/doi/10.1515/9783111508788-009/html.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. "Named Entity Recognition and Classification in Historical Documents: A Survey." *ACM Computing Surveys* 56, no. 2 (February 29, 2024): 1–47. https://doi.org/10.1145/3604931.

Erdmann, Alexander, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Bèatrice Joyeux-Prunel, and Marie-Catherine de Marneffe.
"Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation." 2017. Reprint, Herodotos-Project, 2019. https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation.

Hawes, G. (2023). 'The Manual'. Data Collection principles and practices. Manto Project.

Kron, Colleen, William L Little, James C Wolfe, Petra Ajaka, Benjamin O Allen, Christopher Brown, Marie-Catherine H de Marneffe, et al. "What's In a Name? Issues in Named Entity Recognition," 2019.

Nouvel, Damien, Maud Ehrmann, and Sophie Rosset. *Named Entities for Computational Linguistics*. London-Hoboken: Wiley Blackwell, 2016.

Romanello, M. and Najem-Meyer, S. (2022). Guidelines for the Annotation of Named Entities in the Domain of Classics (18.03.2022). Zenodo. https://doi.org/10.5281/zenodo.6368101

Shipley, G. (2021). Sun, sea, and sky: on translating directions (and other terms) in the Greek Geographers. https://hdl.handle.net/2381/13286000.v1.

Shipley, G. (2024). "Introduction." In *Geographers of the Ancient World*. Vol. 1. Cambridge University Press.