Guidelines for the Annotation of Classics Knowledge Entities

Matteo Romanello (UNIL), Sven Najem-Meyer (EPFL)

Table of Contents

Version 10.12.2021

1 Introduction

The goal of these guidelines is to define a set of rules for the annotation and linking of *knowledge entities* (KEs) in Classics publications. The notion of *knowledge entity* (Zhang et al. 2021) builds upon and expands the traditional notion of named entity by including any entity of interest in a given domain (e.g. algorithms, datasets and software in Computer Science publications, or proteins in the Biomedical domain). Given such a broader definition, these guidelines capture a wide range of entities spanning from persons and geographical locations (ancient, modern and fictional), to bibliographic references (both to primary and secondary sources) and their structural components. Many of these KEs are, in fact, those that we find listed in back-of-the-book indexes of publications in this domain (e.g. index locorum, index nominum, etc.), as they represent an essential means for scholars to find information within publications.

While named entity annotation guidelines exist for other document types in the Humanities such as newspapers (Rosset, Grouin, Cyril, and Zweigenbaum, Pierre 2011; Ehrmann et al. 2020; Menzel, Zinck, and Petras 2020), literary novels (Bamman, Popat, and Shen 2019; Frontini et al. 2020) and archaeological reports (Brandsen et al. 2020), no specific guidelines existed for Classics publications. The guidelines by Colavizza and Romanello (2017) do cover scientific publications in the field of History but are limited to bibliographic entities, thus excluding more universal entities such as people, locations and organisations.

With these guidelines we put forward a unified approach to the annotation of both traditional and bibliographic named entities, tailored to the specificities of Classics publications, both historical and contemporary. The present guidelines were established in the context of the Ajax Multi-Commentary project, where they are used for annotating knowledge entities in historical classical commentaries. During their development, however, a broader range of Classics publications was considered, including abstracts, books and journal articles, which makes these guidelines amenable to a wider range of publications than just commentaries.

2 Annotation of Named Entity Recognition (NER)

2.1 General annotation rules

- Pre- and post-modifiers of proper names should not be annotated (here in italic):
    - Pre-modifier: “of the *emperor* Augustus”, “*éd.* R. Förster”
    - Post-modifier (including appositions): “the Scamander *river*” or “le Baal *syrien*”. An exception to this rule are epithets/epiclesis (see below)
- Punctuation signs that are not strictly part of an entity should be excluded from annotation (e.g. quote signs, brackets, hyphens, commas, etc.).
    - Some typical cases where a punctuation sign is considered part of the entity are:
        - Punctuation indicating abbreviations (e.g. <pers.author>Hom.</pers.author>)
        - Punctuation used in a reference scope to separate hierarchical levels (e.g. Hom. Il. <scope>1.100-48</scope>). However, punctuation that separates multiple scopes is excluded (e.g. commas).
- Enumerations of entities are annotated as separate entities, even when one of the entities is partial:
    - Berlin-New York → <loc>Berlin</loc>-<loc>New York</loc>

    - J. et J. Ch. Balty → <per>J.</per> et <per>J. Ch. Balty</per>
    - Books 2 and 3 → <scope>Books 2</scope> and <scope>3</scope>
- In the case of metonymic expressions, only their intended meaning should be annotated. However, given the type of documents targeted by these guidelines (i.e. scientific publications), such type of expressions is expected to be quite rare.
- Entities contained within verbatim quotations should be annotated only when the quoted text is written in the same language as the containing document (e.g. a quoted excerpt in Latiin within a scholarly article written in German should not be annotated).
- Nested entities are annotated up to the 1st level of depth (i.e. an entity within an entity)
    - In the case of entities contained within an outer bibliographic entity, such as <primary-full>, <secondary-full> etc. (see section 2.3), the outer entity does not count towards determininig the level of depth (in the example below, the location "Apamée" is found at the first level of depth with respect to its containing work title).

> **Example** W. Van Rengen, *Nouvelles inscriptions grecques et latines*, dans les *actes du Colloque Apamée de Syrie. Bilan des recherches archéologiques 1969- 1971*, Bruxelles, 1972, p. 107
>
> <primary-full><pers.author>W. Van Rengen</pers.author>, <work.seclit>*Nouvelles inscriptions grecques et latines*</work.seclit>, dans les <work.seclit>*actes du Colloque* <loc>*Apamée*</loc> *de* <loc>*Syrie*</loc>*. Bilan des recherches archéologiques* <date>*1969-1971*</date></work.seclit>, <loc>Bruxelles</loc>, <date>1972</date>, <scope>p. 107</scope></primary-full>

2.2 Generic entities

Person

- Definition:
    - Entity referring to a definite individual (be it singular or plural), provided that it contains a proper name. Persons include ancient and modern authors, deities, mythological figures, etc.
- Rules:
    - Collective demonyms such as "the Achaens, the Greeks, etc." are not annotated
    - Titles (e.g. academic titles) should not be included
    - English/German possessives should not be included
    - Name variants such as nicknames and abbreviated names are annotated
        - Especially in publications like commentaries, where economy of page space is a driven criterion, it is not uncommon that names of frequently mentioned mythological characters are abbreviated (Achilles → A.).
    - Epithets/Epiclesis should be included, e.g. "Ajax *Télamonide*", "Hermès *Psychopompe*", "Aion *Plutonios*".
    - Patronymics that appear in isolation are annotated only when they refer unambiguously to a definite individual.
        - For example, "the Atreid" can refer to any of the sons of Atreus, whereas "der Pelide" can only refer to Achilles.

        - Person names may contain a location, especially in the case of ancient people (e.g. Arktinos von Milet, Timomachus of Byzantium, or Themison of Samos) where it is used – in absence of last names – to distinguish homonyms. In such cases the location is included as part of the name and annotated as a nested entity.
            - <pers.author>Arktinos von <loc>Milet</loc></pers.author>
        - Arktinos von Milet, Timomachus of Byzantium : overarching `per` entity, because here the place name is needed to distinguish this very person from other homonyms, thus performing a similar function to a last name “Themison of Cyprus”
- Fine grained entities:
    - pers.author A person should be tagged as an *author* whenever is mentioned in contexts that relate to works he/she has authored. Authors can be ancient, modern and contemporary without distinction.
        - For example, Caesar may be mentioned in some contexts as the author of the *De Bello Gallico* and in others in his capacity of statesman (in this latter case, it should annotated as a <pers.other>).
    - pers.editor A person should be tagged as an *editor* when the mention relates to his/her editorial activity: for example, a person is mentioned as the editor of a critical edition of an ancient text, or is credited for a textual conjecture or variant reading.
    - pers.myth Entity referring to mythological characters or religious entities (e.g. the Sirens, the Cyclopes, etc.).
    - pers.other All other types of persons (e.g. historical figures, scholars, political figures) mentioned without a reference to their authorial or editorial role.
- Examples:
    - The arrival of the <pers.myth>Argonauts</pers.myth>
    - What we find in <pers.author>Homer</pers.author>’s <work.lit>Iliad</work.lit>
    - As the Achaens arrived at <loc>Troy</loc>
    - As argued by Prof. <pers.other>West</pers.other>

Location

- Definition:
    - Entity referring to a “**politically or geographically** defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)” (MUC-6 task definition, quoted from Sonar guidelines)
- Includes:
    - Geo-political locations (cities, countries, colonies)
    - Physical locations (continents, rivers, seas, mountains)
    - Fictional locations (e.g. Olympus, Hades, etc.)
    - Named buildings (temples, museums, libraries). According to the definition given in Wikipedia, “a building, or edifice, is a structure with a roof and walls standing more or less permanently in one place”. Thus, structures like altars are not considered buildings.
- Examples:
    - The <loc>Scamander</loc> river

        - The <loc>Bodleian Library</loc> at <loc>Oxford</loc>[1]
        - <loc>Athens</loc> sent ambassadors to <loc>Sparta</loc>
        - Le <loc>temple de <loc>Baalbek</loc></loc>, le <loc>Parthenon</loc>, the <loc>Temple of <pers.myth>Bacchus</pers.myth></loc>

Organisation

- Definition:
    - “Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure” (ACE guidelines)
- Includes:
    - Names of armies/legions, religious groups, but also modern organisations such as publisher names contained within bibliographic references.
- Examples:
    - La <org>IVe legion Scythique</org>

Date

Dates in Classics publications play an important role. They can specify dating of works, historical events, archaeological artefacts, etc. However, the degree of fuzziness with which these dates are expressed may vary substantially: we find dates that refer to a precise calendar year, as well as time expressions that are more vague and less structured. We follow mostly Brandsen’s guidelines (Brandsen et al. 2020) for the annotation of archaeological publications, which also include annotation of historical periods (e.g. Neolithic).

- Definition:
    - “An absolute date is a date whose position on the calendar can be deduced by the sole information present in the date (or temporal expression), without any context.” (Impresso guidelines)
- Rules:
    - Determiners (*der, die, das, the, il, lo, la, le, les,* etc.) are not annotated as part of the entity, but prepositions are included.
        - "<date>en 180</date>"
    - Time expressions that further characterise an absolute date are included in the annotation (e.g. “vers la fin de 201”)
        - "<date>vers la fin de 201</date>"
        - “<date>Late fourth century BC</date>”
    - In the case of range dates, the entire expression identifying the range should be annotated, including e.g. prepositions
        - <date>From 15 BC to AD 193</date>
- Examples:
    - "<date>dans les premiers mois de 202</date>"
    - "<date>dans la première moitié du ne siècle avant notre ère</date>" (from OCR)
    - In the <date>Neolithic</date> (<date>5550-4200 BCE</date>)

[1]Note that, as capitalization indicates, *library* should be considered here as part of the building’s name rather than as a post-modifier.

Work

- Definition:
  - Entity denoting a human creation, be it intellectual or artistic, that can be referred to by its title.
  - "A work is a distinct intellectual or artistic creation" (FRBR guidelines)
  - "Named entities referring to **titled human creations** are to be classified as works or expressions" (Sonar guidelines)
- Includes:
  - literary works, religious works, editions of papyrological and epigraphical sources (e.g. "IG^{2}", "P.Oxy 1.119"), journals.
- Fine grained entities:
  - work.primlit : work used as a primary source in a given context (*Iliad*)
  - work.fragm : the work is an edition of fragmentary texts (e.g. FHG, FrTrGr)
  - work.inscr : the work is an edition of epigraphic texts (e.g. IG, CIL), i.e. texts preserved on inscriptions.
  - work.pap: the work is an edition of papyrological texts (e.g. P.Oxy., P. Leiden), i.e. texts preserved on papyri.
  - work.seclit : work used as a secondary source (title of journal article, book, etc.).
  - work.journal : work is a scholarly journal (newspapers are excluded)
  - work.other : all other titled works (e.g. music pieces, films, paintings, newspapers, videogames etc.)
- Rules:
  - Sections of works (e.g. the second act of Macbeth) are to be annotated separately as <scope> entities.
  - In cases where a definite articles may or may not be considered as part of the work's title (e.g. "*Les* Metamorphoses" vs. "*les* Metamorphoses"), capitalization of the article should be considered as a decisive clue.
  - Expressions such as "Aeschylean drama", "the Trojan Cycle" or "the Catalogue of Ships" should not be annotated as they are titled works.
- Note:
  - Entities of type <work.fragm>, <work.inscr> and <work.pap> may in some cases be annotated as <work.seclit> in secondary source references. For example, the reference "Egesandro, *FHG* IV, p. 421 fr. 41" refers to fragment of a work by Egesandro as a primary source, whereas in "Fragm. hist, graec. (C. Müller), II, p. 498" the referred object is a specific page of the publication itself – but note here that the abbreviations *FHG* and *Fragm. Hist. Graec.* Here do refer to the same work.

Scope

- Definition:
  - Entity referring to a specific section or portion of a work (e.g. "the second act of Macbeth ", "Hom. *Il.* 1.1-10", "p. 318", "v. 328 f.").
  - A scope may be expressed as a range, and the work to which it refers to may or may not be explicitly mentioned in the context.
  - In the case of works whose textual hierarchy comprises multiple levels (e.g. a book divided into chapters and sections, a poem divided into verses), the *scope* indicates how to navigate the cited work to find the exact portion

being referred to (e.g. “vol. 1 pp. 23”). Punctuation signs (typically dot and comma) are used to separate the references to the various hierarchical levels .

- For example, in “Hom. *Il.* 1.1-10” the scope “1.1-10” points to lines 1-10 of the book 1 of Homer’s *Iliad.*

- Rules:
  - Scopes are annotated only when expressed *explicitly,* i.e. as a precise reference to citable units of the cited work such as books, chapters, sections, lines, etc.). Implicit references, such as “in the *two final books* of the Metamorphoses”, are therefore excluded.
  - Scopes can refer to sections or portions of external sources (both primary and secondary) but also to other sections of the same document being annotated (e.g. in a journal article, the author refers to another page or footnotes of the same article). Both type of scopes should be annotated.
  - In some cases it may be difficult to establish whether a certain expression should be annotated as a single or multiple scopes. In such cases, it is best to annotate the entire expression as a single scope.
    - For example, in “Krüger Gr. 53, 6, A. 3” it is uncertain whether “53”, “6” and “A. 3” indicate three distinct portions of the cited work, or they correspond to different levels of the same work (i.e. chapter 53, section 6, absatz (paragraph) 3). Given this uncertainty, it should be annotated as <scope>53, 6, A. 3</scope>
  - Abbreviations that often anticipates *scope* entities such as “supr.” (for *supra*), “infr.” (for *infra*) should not be annotated.
    - For example, “see *infr.* p. 31” → see *infr.* <scope>p. 31</scope>
  - Subsections of a work that have their own name (e.g. the *Life of Severus* within the *Historia Augusta*) are annotated as part of the scope, as they are not part of the work title.
    - “Hist. Aug. Sev. 3,9” is annotated as <primary-full><work.primlit>Hist. Aug.</work.primlit> <scope>Sev. 3,9</scope></primary-full>.
  - An indication of the type of citable units referred to in the scope may or may not be present, and is often abbreviated (l. for line, p. for page, for col. for column, etc.). If present it should be annotated as part of the scope entity.
    - For example, “<scope>chapter 10</scope>”, “<scope>books 1-10</scope>”, “<scope>lines 156-8</scope>”, “<scope>vv. 1-3</scope>”, “<scope>pp. 512-30</scope>”).
  - In the case of concordances, equivalences between two or more scopes are expressed by means of the equal sign (“ = ”), for example in “915—924=961—973”. In such cases, the “=” should not be annotated as part of the scope: <scope>915—924</scope>=<scope>961—973</scope>
  - In the Alexandrinian way of citing Homeric poems, capitalized Greek letters indicate books of the *Iliad* and lowercase letters indicate books of the *Odyssey.* Thus, “Hom. B 1-10” corresponds to “Hom. *Il.* 1-10”, while “β 1-10” to “Hom. *Od.* 1-10”. In such references we annotate the Greek letter as part of the scope. <pers.author>Hom.</pers.author> <scope>B 1-10</scope>.
- Notes:
  - The Sonar guidelines annotate as *works* what we consider as *scopes.*
- Examples:
  - the <scope>second act</scope> of <work>Macbeth</work>

    - “Hom. B 1-10” → <primary-full><pers.author>Hom.</pers.author> <scope>B 1-10</scope></primary-full>

Object

- Definition:
    - Entity referring to man-made physical objects (i.e. material artefacts) such as manuscripts, archival documents, museum objects (vases coins).
    - *Objects* differ from *works* as they do not point to a titled intellectual or artistic creation (e.g. a critical edition, a literary work, etc.) but rather to the physical object itself. Such objects are usually cited through the identifiers that the holding institution has assigned to them (inventary number, catalogue number, shelf-mark).
- Fine grained entities:
    - object.manuscr the object is a library manuscript or old/rare book
    - object.archive the object is an archival fund, series or document
    - object.museum the object is part of a museum’s holdings (e.g. vase, coin, statue, altar)
- Rules:
    - Manuscripts *sigla* should not be annotated (e.g. “L^2”) as they are not unambiguous referrents. In the context of a critical edition or commentary, manuscripts are usually referred to by means of a so-called *siglum,* namely an abbreviations used for a given manuscript instead of its library shelf-mark (which can vary from edition to edition).
    - Common names of manuscripts should be annotated (e.g. “Homer’s Venetus A”).
    - Generic abbreviations for manuscripts such as MS. and MSS. should not be annotated.

Examples:

L'autel de Vaison est aujourd'hui conservé au Musée des antiquités nationales (Saint-Germain-en-Laye), inv. MAN 11058

L'autel de <loc>Vaison</loc> est aujourd'hui conservé au <loc>Musée des antiquités nationales</loc> (<loc>Saint-Germain-en-Laye</loc>), <object.museum>inv. MAN 11058</object.museum>

monument sculpté de la Villa Albani - Helbig4, n° 3355

monument sculpté de la <loc>Villa Albani</loc> - <object.museum>Helbig4, n° 3355</object.museum>

2.3 Bibliographic entities

In continuity with the guidelines defined by Colavizza and Romanello (2017), we distinguish – at a coarse level – between references to primary sources and to secondary sources. Subsequently, within each type, we make some further distinctions based on the content of references, such as for example whether a reference is complete (full) or whether it is partial and thus logically depends on a previously mentioned reference.

Such bibliographic entities are used to group together entities of other types (typically person, work, scope, date, etc.) that form a structured reference to a primary or a secondary source.

Primary sources

Bibliographic references to primary sources are different from discursive mentions of the same sources. The former are structured typically by means of punctuation, while the latter blend with the flow of the containing sentence. Compare for example: “in the second act of Macbeth” and “cfr. Shakespeare, Macbeth act 2”. Scholarly publications tend to cite primary sources by means of structured references but they can also contain mention them in a more discursive form.

Primary-full

A reference to a primary source is considered complete (*primary-full*) only when it contains minimal necessary details to be properly looked up (i.e. resolved).

The minimal entities that *must* be present in order for a primary reference full to be considered full are: 1) a *person* or *work* entity identifying the cited source (often given in abbreviated forms, e.g. Hom. Il.) and 2) a *scope* entity specifying the precise source passage being referred to.
Rules:

- At a minimum, to be considered full, a primary source reference must contain 1) a <pers.author> or <work.lit> entity *and* 2) a <scope> entity.
- One or more <scope> entities contained in a <primary-full> may be separated by punctuation or coordinating adverbs (e.g. “Hom. *Il.* 1.1-10 and 24.105”).

Examples:

Plut. V. Brut. XXX → literary work cited in a canonical way

```
<primary-full>
   <pers>Plut.</pers> <work>V. Brut.</work> <scope>XXX</scope>
</primary-full>
```

Aristot. de Sens. c. VI 446,22 ed. Bekk. → literary work cited according to a specific edition

```
<primary-full>
   <pers.author>Aristot.</pers.author> <work>de Sens.</work> <scope>c. VI 446,22</scope>
<pers.editor>ed. Bekk.</pers.editor>
</primary-full>
```

Ennius (Enn. ann. fr. 11 Skutsch (1985) → literary work preserved in fragments

```
<pers.aauthor>Ennius</pers.aauthor> (
<primary-full>
   (<pers.author>Enn.</pers.author> <work.frag>ann.</work.frag> <scope>fr. 11</scope>
   <pers.editor>Skutsch</pers.editor> (<pubdate>1985</pubdate>)
</primary-full> )
```

Callimachus, fr. 511 Asper, 465 Pfeiffer → literary work preserved in fragments; the citation provides a concordance of the corresponding fragment numbers in two separate critical editions (Asper’s and Pfeiffer’s).

```
<primary-full>
   <pers.author>Callimachus</pers.author>, <scope>fr. 511</scope> <pers.editor>Asper</pers.editor>,
<scope>465</scope>
   <pers.editor>Pfeiffer</pers.editor>
</primary-full>
```

Libanius, Orat., XLVIII, 14 (éd. R. Förster, III, p. 434) → here we annotate as a primary full, followed by a secondary partial. The passage is cited in a canonical way, but additionally the

reference edition is specified. This is conceptually different from a reference to a fragment which can be cited only with reference to a specific critical edition.

<primary-full><pers.author>Libanius</pers.author>, <work.primlit>Orat.</work.primlit>, <scope>XLVIII, 14</scope></primary-full> <secondary-partial>(éd. <pers.author>R. Förster</pers.author>, <scope>III, p. 434</scope>)</secondary-partial>

7 Hist. Aug.,Seu., 3, 9.
8 Ibid., 3, 6.7 ; 9, 4 ; cf. A. Birley, op. cit., p. 95.

Primary-partial

A <primary-partial> is a bibliographic reference to a primary source which, unlike a <primary-full>, does not contain enough details to be able to resolve it, and typically depends – at a logical level – on a previous reference.

Rules:

- Case of reference scopes separated by coordination (e.g. “8.728–41 and 8.751–66”) → annotate as one single <primary-partial> containing two distinct <scope> entities.

de quo Athenaeus loquitur IV 185

de quo <pers.aauthor>Athenaeus</pers.aauthor> loquitur <primary-partial><scope>IV 185</scope></primary-partial>

to transgress the boundary of fas odii (Stat. Theb. 9.4),[4] as in the case of Statius ’ Tydeus (8.728–41 and 8.751–66) or Eteocles and Polynices.

to transgress the boundary of fas odii (<primary-full><pers.aauthor>Stat.</pers.aauthor> <work.lit>Theb.</work.lit> <scope>9.4</scope></primary-full>),[4] as in the case of <pers.aauthor>Statius</pers.aauthor> ’ <pers.myth>Tydeus</pers.myth> (<primary-partial><scope>8.728–41</scope> and <scope>8.751–66</scope></primary-partial>) or <pers.myth>Eteocles</pers.myth> and <pers.myth>Polynices</pers.myth>.

Hom. *Il.* 3.380, 5.445, 20.325, 20.443, and 21.597.

<primary-full><pers.author>Hom.</pers.author> <work.lit>*Il.*</work.lit> <scope>3.380</scope>, <scope>5.445</scope>, <scope>20.325</scope>, <scope>20.443</scope>, and <scope>21.597</scope></primary-full>.

Secondary sources

Bibliographic references to secondary sources point to a wide range of materials: monographs, journal articles, collective volumes, encyclopaedia entries. In the specific case of Classics, secondary sources include scholarship closely related to primary sources such as critical editions, commentaries, and translations.

Fine grained entities:

- Title: <work.seclit>
- Author: <pers.author>
- Editor: <pers.editor>
- Journal: <work.journal>

- Publication date (year): <date>
- Publication place: <loc>
- Publisher: <org>
- Pagination (page range, typically for chapters or articles): <scope>
- Volume (e.g. for journal articles): <scope> (sometimes including volume and issue of a journal, “12(3)”)

Secondary-full

A <secondary-full> is a bibliographic reference pointing to a monograph. It is considered full when at least the following elements are present: 1) author or editor; 2) title; 3) publication date; 4) publication place. Such references are typically found in the bibliography or in footnotes section (especially in the case of journal articles).

Examples:

F. H. Cramer, Astrology in Roman Laws and Politics, Philadelphie, 1954, pp. 207-214

<secondary-full><pers.author>F. H. Cramer</pers.author>, <work.seclit>Astrology in Roman Laws and Politics</work.seclit>, <loc>Philadelphie</loc>, <date>1954</date>, <scope>pp. 207-214</scope></secondary-full>

Secondary-meta

The entity <secondary-meta> captures references to publications *contained within other publications*, such as journal articles, contributions in collective volumes and encyclopedia articles. One distinctive feature of <secondary-meta> references is that they contain more than one work titles (e.g. title of chapter, title of book; title of article, name of journal). The entity types that can be contained within a <secondary-meta> reference are the same as for a <secondary-full>.

Examples:

J. et J. Ch. Balty, *Apamée de Syrie, archéologie et histoire. I. Des origines à la Tétrarchie*, dans *Aufstieg und Niedergang der römischen Welt*, II, 8 (Berlin-New York, 1977), p. 129, n. 184.

<secondary-meta><pers.author>J.</pers.author> et <pers.author>J. Ch. Balty</pers.author>, <work.seclit>*Apamée de Syrie, archéologie et histoire. I. Des origines à la Tétrarchie*</work.seclit>, dans <work.seclit>*Aufstieg und Niedergang der römischen Welt*</work.seclit>, <scope>II, 8</scope> (<loc>Berlin</loc>-<loc>New York</loc>, <date>1977</date>), <scope>p. 129, n. 184.</scope></secondary-meta>

À peine cité par K. Latte, s.v. Orakel dans la Realencyclopädie, XVIII, 1 (1939), col. 862,

À peine cité par <secondary-meta><pers.author>K. Latte</pers.author>, s.v. <work.seclit>Orakel</work.seclit> dans la <work.seclit>Realencyclopädie</work.seclit>, <scope>XVIII, 1</scope> (<date>1939</date>), <scope>col. 862</scope></secondary-meta>,

Secondary-partial

This type of references, typically found in the body's text or footnotes, usually refer to full citations of the cited secondary sources contained in the same or other sections of the document (bibliography, footnotes).

Rules:

- In series of two or more consecutive references to publications by the same author, the author name may be omitted in all references but the first one. In such cases, references with omitted author's name(s) should be annotated as <secondary-partial>.

Examples:

are also found in Müller (2016, 346-79)

are also found in <secondary-partial>
<pers.author>Müller</pers.author> (<pubdate>2016</pubdate>, <scope>346-79</scope>)
</secondary-partial>

R. Turcan, op. cit., p. 116.

<secondary-partial><pers.author>R. Turcan</pers.author>, <work.seclit>op. cit.</work.seclit>, <scope>p. 116</scope></secondary-partial>.

pour cette seconde visite de Septime Sévère à Apamée, cf. J. et J. Ch. Balty, loc. cit., p .130 et n. 188.[2]

pour cette seconde visite de Septime Sévère à Apamée, cf. <secondary-partial>
<pers.author>J.</pers.author> et <pers.author>J. Ch. Balty</pers.author>, <scope>loc. cit.</scope>, <scope>p .130</scope> et <scope>n. 188</scope>
</secondary-partial>.

A phenomenon that is often found in partial references to secondary literature is the so-called ***ibid.*-style of citations**, which consists of using the Latin abbreviations *Id., Ibid., op. cit.* and *loc. cit.* as placeholders for bibliographic information already stated in previous paragraphs or footnotes with the aim to minimizing repetitions. We annotate these abbreviations by choosing the entity type corresponding to the type of information suppressed and replaced by the abbreviation (see examples here below):

- *Id.* is used when referring another publication by the same author as in the previous reference to avoid repeating the author's name. Thus, it should be annotated as <pers.author>.
- *Ibid.* can be used to cite a difference section of a publication already cited in the previous footnotes, in which case it is followed by a *scope* (e.g. page number) and is annotated as a <work.sectitle>. However, *Ibid.* can also be used to refer to the very same section as in the previous reference (i.e. same publication and same scope); in this case it should be annotated as a <scope>.
- *Op. cit.* is employed to refer to a previously cited publication by providing only indication of its author(s) and of the section cited. Thus, it should be annotated as a <work.sectitle>.

[2] A couple of footnotes earlier, the same publication was already cited "J. et J. Ch. Balty, Apamée de Syrie, archéologie et histoire. I. Des origines à la Tétrarchie, dans Aufstieg und Niedergang der römischen Welt, II, 8 (Berlin-New York, 1977), p. 129, n. 184.". So *loc. cit.* here stands for "p. 129, n. 184".

- *Loc. cit.* is used in conjunction with the author's name to refer to the very same section of a given publication which was cited immediately before, thus it corresponds to a <scope> entity.

3. Annotation of Entity Linking (EL)

Entity Linking is the process of establishing correspondances (links) between a mention of an entity in a document and the record corresponding to that entity within a knowledge base. Our knowledge base of choice is Wikidata, meaning that Wikidata URIs are used to unambiguously identify entity mentions (e.g. Q172725 identifies Ajax Telamonius).

Rules:

- Only top level entities should be linked (i.e. nested entities are excluded from linking)
    - For example, in "<loc>Temple of <pers.myth>Zeus</pers.myth></loc>", the nested entity "Zeus" should not be linked.
- What entity types should not be linked?
    - All entities contained within bibliographic references to primary sources (<secondary-full>, <secondary-partial> and <secondary-meta>)
    - <date> entities
    - <scope> entities
- The Wikidata record describing most precisely a given entity should always be preferred.
    - E.g. in the following sentence "Ajax performs his act in the presence of the Sun", *Sun* is a personification of the natural element (i.e. the god Helios); thus it should be linked to Helios (WID Q134270) rather than to the record for the star sun (WID Q525).

4. Text corrections (segmentation, orthography, OCR)

Annotation of sentence boundaries, follows the following **general principle :** In case of doubt, annotators should favor longer sentences. The examples below illustrate this principle:

- Lemmas are not considered sentences on their own but are to be included in the following sentence (e.g. <Sentence> **314.** The first sentence of lemma 314. </Sentence>).
- Tiny sentences such as "*See Tr. 147."* or "*Vgl. Hom. Il. IV. 18*" should be collated to the previous sentence (e.g. <Sentence> A previous sentence. *See Tr. 147.* </Sentence>).

In the case of sentences in the same lemma separated by a period and an m-dash (or long dash: e.g. *...end of first sentence. — New sentence…*), include the m-dash in the second sentence (e.g. <Sentence>— A sentence beginning with an m-dash. </Sentence>.

Correction of OCR in entity surface forms:

- The annotator should provide the correct transcription for entity surface forms that contain OCR errors, (e.g. <pers.author>Luciane</pers.author> instead of <pers.author>Lucianus</pers.author>).

- If the annotated text span contains a now obsolete spelling, this should not be normalized. For example, “Äſchylos” for Aeschylus should not be transcribed as “Aischylos” with normalisation of the long “s”.

Annotation of hyphenated words:

- Words that are hyphenated due to a line break should be marked by the annotator
- Compound words, which normally would contain a hyphen, should not be annotated (e.g. “half-measure”)

References

Bamman, David, Sejal Popat, and Sheng Shen. 2019. “An Annotated Dataset of Literary Entities.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 2138–44. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1220.

Brandsen, Alex, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. “Creating a Dataset for Named Entity Recognition in the Archaeology Domain.” In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4573–77. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.562.

Colavizza, Giovanni, and Matteo Romanello. 2017. “Annotated References in the Historiography on Venice: 19th–21st Centuries.” *Journal of Open Humanities Data* 3 (0): 2. https://doi.org/10.5334/johd.9.

Ehrmann, Watter, Romanello, Clematide, and Flückiger. 2020. “Impresso Named Entity Annotation Guidelines,” January. https://doi.org/10.5281/zenodo.3604227.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. “Named Entity Recognition for Distant Reading in ELTeC.” In *CLARIN Annual Conference 2020*. Virtual Event, France. https://hal.archives-ouvertes.fr/hal-03160438.

Menzel, Sina, Josefine Zinck, and Vivien Petras. 2020. “Guidelines for Full Text Annotations in the SoNAR (IDH) Corpus.” https://github.com/qurator-spk/neat/blob/master/Annotation_Guidelines.pdf.

Rosset, Sophie, Grouin, Cyril, and Zweigenbaum, Pierre. 2011. “Entités Nommées Structurées : Guide d’annotation Quaero.” 2011–04. Orsay, France: LIMSI-CNRS.

Zhang, Chengzhi, Philipp Mayr, Wei Lu, and Yi Zhang. 2021. “Extraction and Evaluation of Knowledge Entities from Scientific Documents.” *Journal of Data and Information Science* 6 (3): 1–5. https://doi.org/10.2478/jdis-2021-0025.