Guidelines for the *Vasari* Named Entity Recognition, Entity Linking and Artwork Detection Evaluation Dataset

Preamble

This guidelines are based on the document *Annotation Guidelines for Named Entity Recognition, Entity Linking and Stance Detection* (Hamdi et al., 2021a) for the NewsEye multilingual dataset of Historical Newspapers (Hamdi et al., 2021b). Scope of these guidelines is the annotation of Named Entities, i.e. linguistic units such as proper names or definite descriptions having the status of a proper name.

1. Entity Types

For the Named Entity Recognition and Entity Linking dataset, the four entity types from CONLL were reused.

Entity	Description
PER	Humans
ORG	Organizations, political and religious groups
LOC	Natural places, buildings and facilities
MISC	Miscellaneous entities, such as cultural themes, historical events, mythical characters.

An additional type, i.e. WORK was added in order to annotate artwork titles and definite descriptions of artworks for the Artwork Detection Task. This task is separated from Named Entity Recognition and Linking task since artworks can be referenced in text not only via official titles but also through complex nominal syntagms which have a descriptive function, as in *altarpiece of S. Maria Novella made by Ghirlandaio's brothers*.

2. Named Entity Mentions Lexical Characteristics

a. Nature: Named Entity mentions are often provided in text by specific grammatical features, such as the use of capitalization or quotation marks (e.g., for referencing titled works). Named Entity mentions given via noun phrases, i.e. definite descriptions, have a less clear status than proper names. For the sake of clarity, definite descriptions are considered as Named Entities if they reference a specific, unrepeatable entity, event or concept. For example, while "siege of Florence" or "house of Cosimo de' Medici" are annotated, "decorated room" will not be annotated since it has not a sufficient degree of specificity.

b. Boundaries: A named entity can be the head of several nominal syntagms, but not all of them are annotated.
Surface forms exclude:

- subordinate clauses
- incidental clauses

- appositives
- determiners, unless they are included in the proper name (*The Arabian Nights*)

Surface forms include:

- pre-modifiers: personal titles (*Saint*, *Duke*, *Pope*), numerals (*twelve apostles*, *two thieves*), epithets (*elder Lorenzo Medici*), artwork types (*statue of Duke Lorenzo*).
- post-modifiers: regnal numbers, epithets (*S. John the Baptist*), toponymic (*Leonardo da Vinci*), patronymics (Leonardo di Lodovico Buonarroti)

c. Coordination: Named entities coordinated based on a coordinate conjunction or a common descriptor are annotated separately.

- <per>Adam<per> and <per>Eve<per>

- <per>Saints Peter<per> and <per>Paul<per>

d. Nesting and special constructions: Nested entities are annotated for the four CONLL entity types, with a limit of depth 1. Nested Entities are annotated with a greedy approach, i.e. the maximum number of nested entities should be annotated.

- <loc>Chamber of
  <per>Our Lady<per>
  at
  <loc>Loreto<loc>
  <loc>

- <per>Giovanni da
  <loc>Santo Stefano a Ponte<loc>
  of
  <loc>Florence<loc>
  <per>

3. Artwork Mentions Lexical Characteristics

a. Nature: With Artwork Mentions we refer to titles of artistic endeavors or definite descriptions of man-made objects which are considered to be endowed with a special aesthetic value and unicity. Since this definition relies on a specific cultural frame, i.e. the Western culture frame, we rely on the annotator's sense of relevance when annotating a specific reference to a cultural heritage object. Moreover, the end goal of extracting artistic information should always be kept in mind.

b. Boundaries: Surface form of artworks can be hard to define since they may define several attributes of an object and not all of them are to be considered.
Surface forms of artworks do not include:

- determiners
- numerals
- adjectives describing shape or size

Surface forms of artworks include:

- the artifact type (*panel*, *portrait*)
- the reference to its subject or depicted content (*model of Hercules killing Cacus*)
- the reference to its author (*David of Donatello*), place (*altar of S. Jacopo*), material and technique (*Crucifixion in wood*).

4. Entity disambiguation and NIL entities

Named Entities and Artwork Mentions are disambiguated by using the Wikidata Knowledge Graph (www.wikidata.org). For our disambiguation, the goal of our annotation is to link the entity to the most specific resource of the KB. However, this is not always possible, especially when certain entities are out of the reference KB. In this case, entities are labeled as NIL (Out-of-KG).

5. Special cases

a. Multiple true links for a single surface form: In some cases, more than one link can be possible for the same surface form. This is especially true for artworks which have the same title as their depicted theme, as in *Ecce Homo*. In this case, two overlapping annotations are provided, one for the subject (MISC) and one for the artwork (WORK), with two different KB identifiers.

b. Ambiguous Named Entities: Since we are dealing with excerpts from books, often historical persons can be referenced only with their first name, since the original full name may be given in other parts of the book or is implicit. This causes a certain ambiguity which often requires the annotator to take in consideration the context to understand what is the correct link for a surface form. In this case, we decided to annotate first names too, however we decided to not penalize a model which is not capable of disambiguating them correctly by labeling them both as NIL and with a corresponding Wikidata identifier, if available.
The same situation occurs for capitalized words in Vasari's book which do not refer to a specific individual or entity, such as "Pope", "Saint", "King", etc. For these entities, we decided to provide two possible links:

- a *very* specific link, disambiguating the entity with the specific individual which is mentioned
- an *appropriate* link, which is not the entity truly mentioned by the author but rather represents the concept conveyed by the word.

Example:

- The Sistine Chapel was commissioned by the [Pope]$_{\{\text{Pope Sixtus IV, Pope}\}}$