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). These guidelines further extend the scope of the referenced document by taking in consideration the concept of artwork detection task and by introducing some special cases related to named entity mentions in the *Lives of The Artists* book.

1. Entity Types (NER)

This dataset reuses the four entity types from CONLL-2003 for the Named Entity Recognition task. In specific, these entity types are:

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

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
- non-essential 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

a. Nature: The artwork detection task is a sequence labeling task which aims to identify *proper names* of artistic objects. This introduces a distinction between the NER task since definite descriptions are not taken in consideration. Artworks are furthermore distinguished from built works (churches, palaces, etc.) to avoid annotation redundancy, since these entities are already annotated as Named Entities and labeled as `LOC`.

Overall, for the annotation of artwork entities, the annotators took in consideration the scope of the extraction task, which is the recognition and linking of entities related to a specific art-historical domain, i.e. visual arts from the late Gothic to the Mannerist Era. This art-historical period was characterized by the predominance of representational arts, i.e. involving non-abstract subjects. Therefore, it was decided to label as `WORK` all Named Entities which are referenced as the primary subject of a visual work (either a statue or a painting).

- A pattern of four pictures, painting therein the <work>Nativity of Christ<work>, the <work>Flight into Egypt<work>, and the <work>Massacre of the Innocents<work>.

b. Boundaries: Titles of artworks can be hard to define. Since most artistic works in *Lives of the Artists* are presented through an indirect description, i.e. *ekphrasis*, as a rule of thumb we annotated as `WORK` all Named Entities which occur inside an ekphrasis as the primary subject of a visual work. This means that surface forms of artworks consist of a sequence of single or multiple named entities which are *represented* inside an artifact, along with clauses which define the *state* in which they are represented.

- An altar-picture of <work>Christ taken down from the Cross, with the Thieves fixed on their Crosses, and the Madonna in a swoon<work>.
- Model of a <work>nude David who was holding Goliath under him and was cutting off his head<work>.

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]$_{\{\text{Pope Sixtus IV, Pope}\}}$