

Sierra-Múnera, Alejandro; Le, Linh; Demartini, Gianluca; Krestel, Ralf

Article — Published Version

MELArt: A Multimodal Entity Linking Dataset for Art

Transactions on Graph Data and Knowledge (TGKD)

Suggested Citation: Sierra-Múnera, Alejandro; Le, Linh; Demartini, Gianluca; Krestel, Ralf (2024) :
MELArt: A Multimodal Entity Linking Dataset for Art, Transactions on Graph Data and Knowledge
(TGKD), ISSN 2942-7517, Schloss Dagstuhl - Leibniz Center for Informatics, Dagstuhl, Vol. 2, Iss. 2
(Article No.): 8,
<https://doi.org/10.4230/TGDK.2.2.8>

This Version is available at:
<http://hdl.handle.net/11108/655>

Kontakt/Contact

ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics
Düsternbrooker Weg 120
24105 Kiel (Germany)
E-Mail: info@zbw.eu
<https://www.zbw.eu/de/ueber-uns/profil-der-zbw/veroeffentlichungen-zbw>

Standard-Nutzungsbedingungen:

Dieses Dokument darf zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. Sie dürfen dieses Dokument nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Sofern für das Dokument eine Open-Content-Lizenz verwendet wurde, so gelten abweichend von diesen Nutzungsbedingungen die in der Lizenz gewährten Nutzungsrechte.

Terms of use:

This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.



<https://creativecommons.org/licenses/by/4.0/>

MELArt: A Multimodal Entity Linking Dataset for Art

Alejandro Sierra-Múnera ✉🏠^{ID}

Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

Linh Le ✉🏠^{ID}

The University of Queensland, Brisbane, Australia

Gianluca Demartini ✉🏠^{ID}

The University of Queensland, Brisbane, Australia

Ralf Krestel ✉🏠^{ID}

ZBW – Leibniz Information Centre for Economics, Kiel, Germany

Kiel University, Kiel, Germany

Abstract

Traditional named entity linking (NEL) tools have largely employed a general-domain approach, spanning across various entity types such as persons, organizations, locations, and events in a multitude of contexts. While multimodal entity linking datasets exist (e.g., disambiguation of person names with the help of photographs), there is a need to develop domain-specific resources that represent the unique challenges present in domains like cultural heritage (e.g., stylistic changes through time, diversity of social and political context). To address this gap, our work presents a novel multimodal entity linking benchmark dataset for the art domain together with a comprehensive experimental evaluation of existing NEL methods on this new dataset. The dataset encapsulates various entities unique to the art domain. During the dataset creation

process, we also adopt manual human evaluation, providing high-quality labels for our dataset. We introduce an automated process that facilitates the generation of this art dataset, harnessing data from multiple sources (Artpedia, Wikidata and Wikimedia Commons) to ensure its reliability and comprehensiveness. Furthermore, our paper delineates best practices for the integration of art datasets, and presents a detailed performance analysis of general-domain entity linking systems, when applied to domain-specific datasets. Through our research, we aim to address the lack of datasets for NEL in the art domain, providing resources for the development of new, more nuanced, and contextually rich entity linking methods in the realm of art and cultural heritage.

2012 ACM Subject Classification Computing methodologies → Information extraction

Keywords and phrases A Multimodal Entity Linking Dataset, Named Entity Linking, Art Domain, Wikidata, Wikimedia, Artpedia

Digital Object Identifier 10.4230/TGDK.2.2.8

Category Resource Paper

Supplementary Material The source code for generating MELArt is published on Github under an MIT license and the code for the experiments described in Section 4.3 is also published on Github. The annotations included in MELArt and the candidates' information except for the image files, are available on the UQ Research Data Manager under an CC-BY license. The image files can be extracted from Wikimedia using a file contained in the same repository and a script described in the Github repository.

Dataset (Dataset): <https://doi.org/10.48610/8a1ccdf> [10]

Software (Dataset generation code): <https://github.com/HPI-Information-Systems/MELArt> [11]
archived at `swb:1:dir:ec4380448f4087c011040d0e3dca7832baa11182`

Software (Experiments code): https://github.com/HPI-Information-Systems/MELArt_experiments [12]
archived at `swb:1:dir:203f2a69c5bc9064db3873a0160ca52f62095c25`



© Alejandro Sierra-Múnera, Linh Le, Gianluca Demartini, and Ralf Krestel;
licensed under Creative Commons License CC-BY 4.0

Transactions on Graph Data and Knowledge, Vol. 2, Issue 2, Article No. 8, pp. 8:1–8:22



Transactions on Graph Data and Knowledge

TGDK Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Funding The article is based upon work conducted as part of the Australia–Germany Joint Research Cooperation Scheme (Universities Australia – DAAD). Project number 57600378.

Alejandro Sierra-Múnera: Funded by the HPI Research School on Data Science and Engineering

Received 2024-07-02 **Accepted** 2024-11-15 **Published** 2024-12-18

Editors Aidan Hogan, Ian Horrocks, Andreas Hotho, Lalana Kagal, and Uli Sattler

Special Issue Resources for Graph Data and Knowledge

1 Introduction

The art world, rich in historical and cultural value, is facing a noticeable challenge in the application of computational techniques due to the limited availability of comprehensive datasets for machine learning tasks, despite the existence of some notable datasets, like, for example, The Metropolitan Museum of Art’s Open Access dataset [17], which offers over 397,121 images of public-domain artworks with rich metadata including 224,208 classes. Similarly, the Rijksmuseum dataset includes extensive collections of Dutch art, pivotal for studies in European art history. WikiArt [7] and Google’s Art & Culture datasets¹ provide broader scopes, encompassing diverse global art pieces. Additionally, Artemis [1] stands out as a unique dataset focusing on the emotional responses to art, offering a different dimension for analysis. Artpedia [13], further enriches the available resources for computational art research by associating paintings with corresponding visual and contextual sentences from Wikipedia. IICONGRAPH [9], defines a knowledge graph focused on iconographic and iconological statements for the Italian cultural-heritage landscape.

Despite these advancements, none of these resources are tailored for the task of entity linking or entity disambiguation, in which connections between art subject mentions and their corresponding entities in a knowledge graph can be automatically discovered.

Artpedia, as a comprehensive art resource, offers vast potential for developing an art entity linking dataset. It features a diverse collection of artworks, detailed metadata, high-quality images, and rich contextual information, making it a reliable source for art research [13]. The manual selection of text related to the artworks in Artpedia is its strength, enhanced by the manual classification between *visual* (describing what is depicted in the artwork) and *contextual* (describing other aspects of the artwork) sentences. The detailed descriptions and contextual information provided by Artpedia are vital for textual analysis, which, when combined with visual data, can significantly enhance entity recognition and linking accuracy. Utilizing Artpedia’s resources to develop an entity linking dataset would not only leverage the reliability of Artpedia’s textual descriptions, but also offer new insights into the complexity between visual elements and contextual information in artworks, advancing both the technological and cultural understanding of art.

Wikidata, although being a general purpose knowledge graph, contains specific relations that connect artworks with their corresponding subjects, and, compared to other domain-specific resources, covers more artworks with a good density of subjects per artwork [3]. This coverage, plus the tight connection between Artpedia, Wikipedia and Wikidata, suggests the potential to combine these resources and integrate textual and structured connections between artworks and subjects.

In this paper, we present a new dataset for multimodal (i.e., containing textual mentions and images) named entity linking in art that addresses the limitations of current datasets by: i) integrating Artpedia’s detailed textual descriptions of artworks to enhance the entity recognition

¹ <https://artsandculture.google.com/>

and linking; and ii) meticulously integrating Artpedia with Wikidata; iii) retrieving structured information about paintings from Wikidata; vi) focusing on named entities in the depicted art and expanding their reference labels; vii) implementing named entity recognition and linking using these expanded labels.

2 Entity Linking Challenges in the Art Domain

Entity linking, involves the identification, disambiguation, and connection of entities mentioned in text, to their corresponding entities in a knowledge base. In the specific realm of entity linking, the art field presents unique challenges. In the art domain, it is especially important to not only consider text, but also the images associated with the artworks.

However, the scarcity of datasets tailored to entity linking in artwork is a significant barrier to developing effective techniques. This domain is filled with subjective interpretations and ambiguities, complicating the task of entity linking [5]. For example, a single symbol or motif might have varying meanings in different cultures and historical periods, requiring extensive contextual knowledge for precise entity identification. The inconsistency in image quality, influenced by the age of the artwork or the materials used in its creation, often results in detail or color fidelity loss. This variability adds complexity to the processes of entity recognition and linking, needing algorithms to handle these idiosyncrasies. Therefore, developing efficient entity linking methods in the art domain involves addressing these challenges and highlights the need for custom art-related entity linking datasets, particularly those that are multimodal. Multimodal entity linking datasets for the general domain have been proposed. However, in contrast to datasets like WikiMEL [14], TwitterMEL [2] and WikiDiverse [15], where the images are photographs and the entities are typically persons, for the majority of art subjects the visual representations in MELArt, the dataset we propose in this work, come from paintings. These artistic representations of the subjects come with a high variety of styles and contexts, making it challenging for an automated model to deal with. We adhere to the multimodal entity linking task defined in these datasets, treating the input as a sentence with a mention accompanied by an image, and the objective is to link the correct entity to the mention. An example of the task, specific to the art domain, can be found in Figure 1. Here the input is the sentence “Mary, sitting on a throne . . .” with the corresponding mention “Mary”, with the additional image from the painting. The output should be a score or a ranking of candidate entities, like the candidates on the right in Figure 1, in which the ground truth entity (Q345 in this example) should be at the top of the ranking. Including the image of the painting and potential images for the candidate entities is essential for the art domain, given the visual nature of artwork.

The art domain presents distinctive challenges due to the rich historical context, variability in artistic descriptions, and the interplay of various figures and symbols. We want to illustrate the unique challenges of entity linking within the art domain through specific example sentences with links to Wikidata entities.

Complex Scene Composition. The example sentence “Here the Virgin lifts the veil over the sleeping Child, who is turned toward the audience, with her other arm around the young John, who has a reed across his shoulder” contains the NEL challenge of identifying and linking multiple figures present in a single artwork. The entities involved in this example are the Virgin Mary (Q345), the Christ Child (Q942467), and the Child Saint John (Q1698874). The NEL task complexity arises in distinguishing each figure and associating them with their respective symbolic attributes, like the veil and the reed.

8:4 MELArt: A Multimodal Entity Linking Dataset for Art

Annunciation with St. Margaret and St. Ansanus (Q979440)
painting by Simone Martini and Lippo Memmi



"Mary, sitting on a throne, is portrayed at the moment that she is startled out of her reading, reacting with a graceful and composed reluctance, looking with surprise at the celestial messenger."



Virgin Mary (Q345)

human biblical figure, human. mother of Jesus



Mary, Queen of Scots (Q131412)

human. Queen of Scotland from 1542 to 1567



Mary Magdalene (Q63070)

human biblical figure, human. follower of Jesus

■ **Figure 1** Multimodal entity linking example in the art domain. On the right, we show 3 potential entity candidates for the mention *Mary* marking *Virgin Mary (Q345)* as the correct entity for the artwork on the left. The text corresponds to a *visual sentence* from the Artpedia dataset.

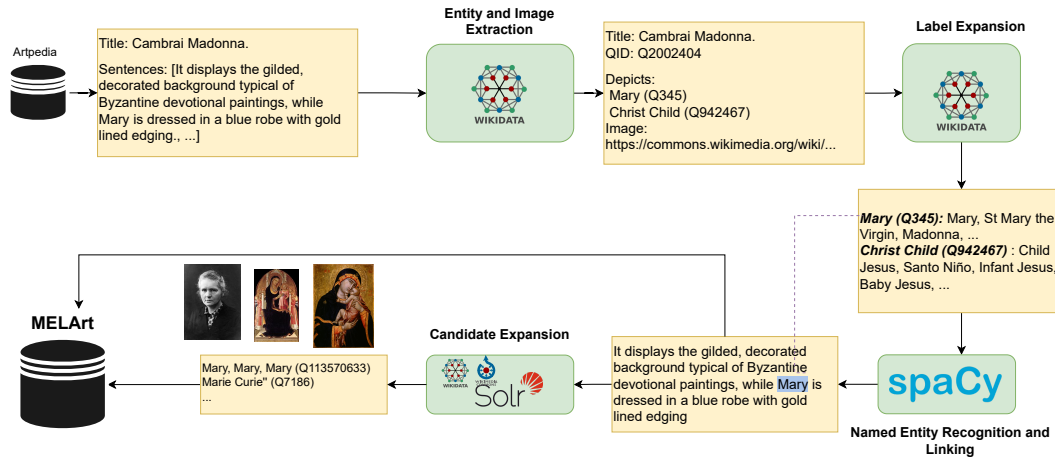
High Disambiguation. The sentence “Queen Charlotte dropped behind Montagne due to the loss of this mast and thus failed to capture her, and led to criticism of the painting” presents a sophisticated entity linking challenge. The primary task is to accurately identify and link “Queen Charlotte” to the entity HMS Queen Charlotte (the ship - Q680379), a task complicated by the potential misinterpretation of “Queen Charlotte” as a person entity, especially given the use of the pronoun “her” in the sentence. This situation underscores a notable aspect of entity disambiguation, where the capacity for multimodal entity linking to discern between human and non-human entities becomes crucial. Such a capability is particularly significant given that many existing datasets exhibit a pronounced bias towards the presence of person entities, highlighting the need for more nuanced and context-aware approaches in entity linking methods for the art domain.

Artistic Historical Linkage. The sentence “The central panel shows the Crucifixion of Christ with John the Apostle and the Virgin Mary” necessitates a precise entity linking process, wherein the name mention “Christ” must be correctly linked to the entity Jesus (Q302) rather than Christ Child (Q942467). The accurate identification of this historical event hinges on the detection of the Cross in the image, which serves as a critical visual cue signifying the Crucifixion. This visual element is essential for distinguishing between the adult Jesus at the Crucifixion and other representations of Jesus, such as the Christ Child. The presence of other key figures in the narrative, such as John the Apostle and the Virgin Mary, further contextualizes the scene, reinforcing the correct linkage to the Crucifixion event. This example highlights the importance of contextual and iconographic cues in historical art for accurate entity linking, especially in complex religious narratives where different phases of a central figure’s life are depicted.

Through these examples, we can see how the art domain poses unique challenges for entity linking, including the need for disambiguation, dealing with variable descriptions, understanding the historical and artistic context, distinguishing between multiple entities, and recognizing artistic authorship. Each aspect addresses the multifaceted nature of entity linking in the realm of art history and criticism.

3 Dataset Construction

Our art entity linking dataset construction method, as depicted in Fig. 2, incorporates a multi-step process that is in alignment with the critical components of the dataset. The following details each step in the process:



■ **Figure 2** The MELArt automatic construction process.

- **Extraction of Painting Title and Visual/Contextual Sentences:** The Artpedia title of the painting, such as “Portrait of Mlle Rachel”, along with contextual sentences like “Portrait of Mlle Rachel is an oil painting on millboard”, are extracted from the Artpedia dataset.
- **Entity and Image Extraction (Multiple-art sources integration):** This step involves querying the unique identifier of the painting from Wikidata, like, for example, “Q979440” (see Fig. 1). Additionally, the image of the painting is retrieved from Wikimedia. This step is discussed in Section 3.1.
- **Label Expansion:** This phase entails expanding the set of surface forms used to refer to the depicted named entities, taking advantage of Wikidata’s alternative labels. This step is discussed in Section 3.2.
- **Named Entity Recognition and Linking:** This process identifies occurrences of names within texts. It utilizes the expanded set of labels to cover the different ways used to refer to the depicted entities. This step is discussed in Section 3.3
- **Candidate Expansion:** This step creates an extensive set of Wikidata entities that, besides the depicted entities, could be referenced from the text. This is a crucial step to make our dataset valuable for the evaluation of NEL models. This step is discussed in Section 3.4.

3.1 Entity and Image Extraction

Given the need to ensure precise matching and integration of data across data sources, to efficiently merge Artpedia with the two other online databases, Wikidata and Wikimedia Commons, a structured approach is necessary. Our process can be systematically broken down to:

1. **Locating the artwork’s Entity-ID by Title:** Identify and locate the specific entity in Wikidata using its Artpedia title (which is equivalent to its Wikipedia article and unique inside Artpedia). This step is implemented by querying Wikidata’s links to Wikipedia articles using the following SPARQL query²:

² All the Wikidata extraction is performed on a dump downloaded on 2024-09-13 and generated on 2024-09-06T23:44:15Z installed on QLever [4] as SPARQL engine.

```
PREFIX schema: <http://schema.org/>
SELECT ?wikipedia_title ?wikidata_id WHERE {{
  VALUES ?wikipedia_title {"@title"@en}.
  ?wikipedia_id schema:name ?wikipedia_title.
  ?wikipedia_id schema:about ?wikidata_id.
  ?wikipedia_id schema:isPartOf <https://en.wikipedia.org/>.
}}
```

where *@title* is the title from Artpedia. In this process we found that for a subset of 20 paintings, the match was not possible. The two main reasons for a missing match were: i) deleted Wikipedia articles like “Charles II in armor” and ii) Wikipedia articles that became disambiguation pages like “The Three Musicians (painting)”. For those cases we manually defined the matching between Artpedia’s title and Wikidata.

2. Extract from Wikidata, using SPARQL, the following predicates for each artwork:
 - P18: *image*. We extract the URL of the first image for each painting that is part of Wikimedia Commons.
 - P180: *depicts*. This corresponds to each “entity visually depicted in an image, literarily described in a work, or otherwise incorporated into an audiovisual or other medium”.

3.2 Label Expansion

In this step, we take advantage of the rich set of alternative labels defined in Wikidata entities to consider the multiple surface forms that might be present in Artpedia artwork description sentences to refer to depicted entities. To do so, we extract Wikidata’s alternative labels for the depicted entities. For each of the depicted entities, we list all the available alternative labels. For example, the entity *Virgin Mary (Q345)* can be referred to by multiple surface forms, which include among others: *Our Lady*, *The Virgin Mary*, *Madonna*, and *Holy Virgin*.

At the end of this process, the total number of labels for depicted named entities becomes 12,966, averaging 2.9 labels per depicted entity.

3.3 Named Entity Recognition and Linking

In this stage, we create a rich set of mention-entity pairs by finding the depicted entity labels inside Artpedia’s sentences. This pair construction is a critical step, as it forms the backbone of the MELArt dataset, linking the entities mentioned in the artwork descriptions to their corresponding entities within a structured and accessible knowledge base. This structure is critical for facilitating research and applications in art and cultural heritage data analysis. The steps we use are the following:

1. Select named entities. The depicted entities identified in the previous step are carefully filtered to preserve named entities and remove general entities. Given the set of all the entities depicted in an artwork, we decide to only keep named entities as iconographic statements. In our process, this is important because automatically matching non-named entities might create false positives since their surface forms are common words. Similar to [9], to determine if an entity is a named entity (i.e., an iconographic statement), we extract the English labels, and check if any of them contain an uppercase character. In this way we avoid entities like *horse* or *hand* and keep entities like *Christ* or *Paris*. An initial evaluation of using Wikidata entity types (i.e., using *instance_of* predicates) as a filter for named entities, with a manually defined set of classes like *human* and *city*, showed smaller coverage of entities as compared to the capitalization heuristic. An inspection of the missed entities showed that the selection of the set

■ **Table 1** Compilation of the Name Mention “Gustave Courbet” and different entities sharing this label.

Label	Description	Instance of
Gustave Courbet	French painter (1819-1877)	Human
Gustave Courbet	Exhibition Fondation Beyeler	Art Exhibition
Gustave Courbet	2007–2008 Exhibition at Metropolitan Museum	Art Exhibition
Gustave Courbet	Print in the National Gallery of Art (NGA 162984)	Print
Gustave Courbet	(Temporary) Art exhibition	Art Exhibition
Gustave-Courbet-Straße	Street in Mitte district, Berlin, Germany	Street
Gustave-Courbet-Straße	Street in Taucha, Saxony, German	Street
Gustave Courbet’s Meeting: A Portrait of the Artist as a Wandering Jew	Journal article; published in 1967	Academic journal article
Gustave Courbet’s “The Sleepers”. The Lesbian Image in Nineteenth-Century French Art and Literature	Article	Scholarly Article
rue Gustave-Courbet	Street in Paris, France	Street
Allée Gustave-Courbet	Alley of Montreuil in Seine-Saint-Denis, France	Allée

of classes needed to cover all the depicted named entities in the dataset is a manually intensive and error-prone task. We thus refrain from manually defining and maintaining such an entity type set and instead follow the more automated process that looks at letter capitalization.

2. Match the labels to the artwork sentences. For each of the paintings, we match any of the labels of the depicted named entities to the visual and contextual sentences from Artpedia, using the Spacy’s Phrase Matcher³. The result is a set of spans in the sentences corresponding to the depicted entities. Given that the list of possible entities is limited by the depicted named entities, this approach focuses on high matching precision. The matching is case-sensitive, to avoid false positives, especially with names like *Our Lady* which are composed of common words.
3. Filter nested mentions. In the case of overlapping matches, where one match is completely contained in another match, we select the longer and more specific entity mention. For instance, if the entity *Madonna and Child* is recognized as an entity, but in the same span *Madonna* is also recognized, then the latter will not be considered.

3.4 Candidate Expansion

In this stage, given the recognized and linked mentions, we build an extensive set of Wikidata entities that can be potentially linked to the depicted entities. We base our expansion on the assumption that entities sharing labels are challenging to disambiguate, with a limited context.

³ <https://spacy.io/api/phrasematcher>

Table 1 illustrates various entity candidates for the name “Gustave Courbet”. It exemplifies the multifaceted challenges inherent in entity disambiguation tasks within art datasets. The table shows the wide variety of meanings that can be connected to a single name, highlighting how one term can have many different interpretations or associations. The labels in the table range from the person “Gustave Courbet” himself to various exhibitions themed around him, and even streets named after him. This example shows how widely a single name can be interpreted, pointing to different entities or concepts each time it is mentioned. The descriptions confer granularity, distinguishing the historical individual from specific art expositions and location entities. Crucially, the “Instance of” category elucidates the ontological nature of each referent, spanning classifications like “Temporary Exhibition”, “Print”, and “Scholarly Article.” This typological diversity underscores the need for an advanced semantic analysis approach, incorporating label and feature extraction (from both text and image) to ensure precision in the mapping of nominal references to their corresponding entities within a knowledge base such as Wikidata. The steps we use are the following:

1. Wikidata labels full-text index. We first index all the English entity labels from Wikidata into an Apache Solr⁴ core. This system is configured to index the text of the labels using an English analyzer composed of tokenization, stop word removal, possessive filter and stemmer. In addition to the labels, we include the number of Wikipedia articles for each entity to represent the popularity of the entity.
2. Expand the candidates list using the mentions’ text. Given the set of matched mentions in the sentences, we expand the set of candidate matching entities for those mentions using the indexed Wikidata labels. For each mention (e.g., “Mary”) we perform two queries to the Solr index. First, we find the 25 most similar labels to the mention, based on the BM25 algorithm ($k_1=1.2, b=0.75$). This returns candidates with very similar labels like “Mary, Mary, Mary” (Q113570633). Additionally, we perform the same search but sorting the results according to the number of Wikipedia links in descending order and choosing the first 25 results. This query finds entities like “Marie Curie” (Q7186) with only a partial label match, but with a high popularity. We then include these two sets of candidates and the ground truth entity to the universe of MELArt candidates. This final set contains 53,901 entities.
3. Extract details from the candidates. For each of the candidate entities, we use SPARQL to extract the following predicates.
 - P18: *image*. For each candidate entity, we extract all Wikimedia Commons images, except those that correspond to the Artpedia artwork images. We filter those because for some candidate entities, the related images already correspond to the paintings in which they have been depicted. If we were to keep them, that would make the entity linking task trivial in those cases. For instance, the painting *Ophelia*(Q21192340) by *John William Waterhouse* depicts the character *Ophelia*(Q1800888) from Hamlet, but the image associated with the Wikidata entity of the character is Waterhouse’s painting.
 - P31: *instance of*. We extract the classes the entity belongs to, and their main labels.
 - We extract Wikidata’s English description of the entity.
4. Download images. Together with the painting image links, and all the candidate images, the image files are downloaded from Wikimedia Commons or Wikipedia in a few cases. Wikimedia Commons links are given priority in case there are multiple images for a painting.

⁴ <https://solr.apache.org/>

4 Evaluation and Results

We perform two types of evaluation of the quality of the MELArt dataset. First, we evaluate the reliability of the automatically annotated dataset by employing human annotators who check its correctness. The second involves the use of MELArt with state-of-the-art (SOTA) entity linking models.

4.1 Data Quality Evaluation Using Human Annotations

To evaluate the quality of MELArt, we follow a multistep process with two human annotators that manually complete the entity recognition and linking tasks on a sample of the painting description sentences. The goal of this evaluation is to answer the following question: are the automatically annotated mentions precise and exhaustive? Ideally, all the annotations should correspond to correct entity mentions, and they should be linked to the right Wikidata entities.

We limit the scope of this evaluation to annotations of linked Wikidata entities, thus avoiding the identification of depicted entities not linked in Wikidata, even though there are hints in the text indicating their presence. For instance, for the painting *Rucellai Madonna* (Q948745) and the visual sentence “The painting depicts the Virgin and Child enthroned, surrounded by angels on a gold background.” an annotator could identify *Madonna and Child* (Q9309699) as a good fit for the “Virgin and Child” span, however, according to Wikidata’s triples, this artwork depicts *Virgin Mary* (Q345) and *Christ Child* (Q942467) as independent entities. In such cases, we assume Wikipedia’s links as the ground truth and refrain from adding extra links.

Step 1: Initial joint exploration of a small sample

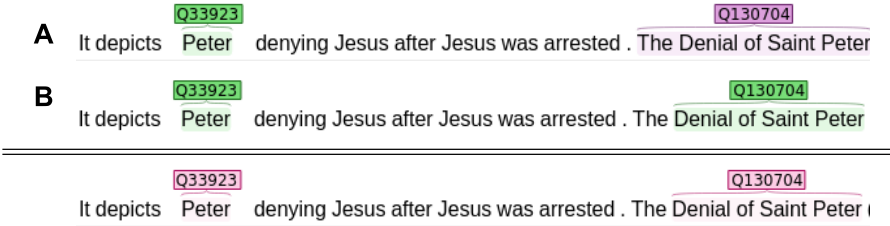
The initial step involves the selection of a sample containing 30 paintings, which was presented to both annotators. The annotators jointly explored the sample, identifying and discussing annotation errors and agreeing on precise definitions to address borderline cases. From this initial data exploration step, the following coding rules were agreed upon:

- In the presence of nested entities, the more specific one must be chosen, and the nested annotations must be removed. For instance, the entity *Madonna and Child* is preferred to annotating *Madonna* if both entities are present in the artwork.
- The article preceding the entities should be included, only if it is capitalized. For instance, “The painting depicts the [Virgin] and ...” should not include the article *the*, while “... from left to right, John the Baptist, [The Virgin Mary] with ...” should include it.
- Only nominal mentions are considered, avoiding pronouns referring to the entities in a different span of text.

Step 2: Independent Annotation

In this step, a bigger random sample of 100 painting descriptions was drawn from the dataset. Annotators worked independently on their assigned tasks to mitigate bias. The annotators did not have access to the annotation made by the other annotator, but they had access to the automatic annotations. The annotators based their decision on their own understandings of the task and the available information about the paintings in Wikipedia and Wikidata.

After this step, we compared the annotations derived from each annotator and computed inter-annotator agreement using Krippendorff’s Alpha. The value of *alpha* after the second step was **0.89**, which can be interpreted as a high level of agreement.



■ **Figure 3** Example of the curation of an annotation disagreement. The top annotations correspond to annotator A and B. The bottom annotation is the curated ground truth as determined by the discussion among annotators.

Step 3: Aggregation and Definition of the Ground Truth

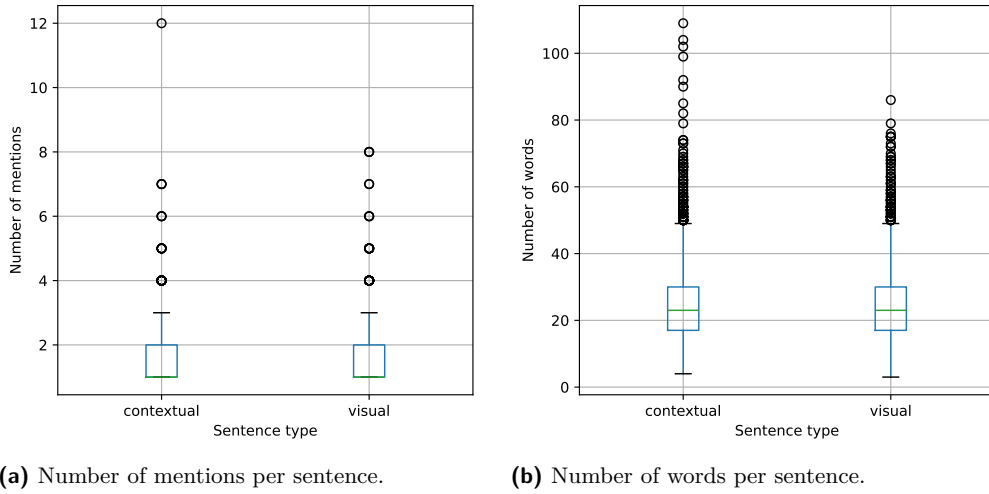
The final step combines the individual annotations into a unified dataset, resolving the few discrepancies through discussion among the human annotators. Both annotators explored the set of 100 documents, using the curation functionalities of the INCEPTION tool, and discussed disagreements until the set of annotations was considered curated and consistent. We consider this curated sample as ground truth entity linking annotations under the scope of the ‘depicted’ predicate of Wikidata. Figure 3 shows an example of the resolution of a disagreement. In this example the disagreement relates to the boundaries of the mention and whether the mention should include the article *The* for the second mention. As per the rule discussed in Section 4.1 the article should be included, but given that it is at the beginning of a sentence, and it is capitalized for that reason, and the name of the entity in Wikidata does not contain the article, the annotators agreed not to include it. In the same example, we show the cases in which named entities are not included in the dataset. Here “Jesus” is an entity mentioned in the text, but not depicted in the painting, thus not part of the multimodal annotations of MELArt.

Step 4: Using Sequence Labeling Metrics to Measure MELArt Quality

In the last step, MELArt annotations are compared against the final version of the ground truth for the sample of 100 paintings. Specifically, we compute precision, recall and F1 for the presence of manually annotated mentions in the automatically generated annotations. To consider a mention as a true positive, the left or right boundaries of the span must match, as well as the assigned entity id.

Human Evaluation Results

The evaluation metrics for MELArt data against the manually curated ground truth focus on the assessment of the dataset’s quality. We observed an F1-score of 0.79 with Precision being notably high (0.90). This reflects the dataset’s high level of reliability in terms of correctly identifying relevant items. The majority of false positives, correspond to mentions not considered as named entities by the human annotators, but included by MELArt heuristics. An example is several mentions to the entity “mother” (Q7560), which contains one label in uppercase. The recall score (0.70), indicates the dataset’s comprehensive nature, encompassing a broad range of relevant items. However, it also shows how automatically matching the labels of Wikidata to natural language does not cover all possible surface forms. This was also observed during the annotation process, in which multiple entities were identified as missing. One of the most common patterns missed by our automatic annotation approach are mentions only containing the first name, last name, or portions of the entity name. For example, the entity *Frida Kahlo* is mentioned in certain cases as



■ **Figure 4** Complexity of sentences per sentence type.

Frida, which, given the context, can be identified by a human annotator as being the artist, but may be missed by the automatic processes we used to generate MELArt. On the other hand, we believe that the annotations contained in MELArt serve the purpose of training and evaluating next-generation approaches for the multimodal entity linking task in the art domain, given its high precision. It is also important to note that the traditional NEL task starts with a set of pre-defined mentions, so NEL-specific models are not directly affected by missing mentions. Ideally, more mentions could be included in the dataset, including pronominal mentions, increasing the difficulty of the linking task. However, imprecise while exhaustive annotations would hinder the evaluation benchmarking potential of the dataset.

In general, the reported metrics suggest that the automatically generated dataset is of high quality, offering a reliable and accurate basis for the training of multimodal NEL models.

The published version of MELArt, used for training and evaluation, includes the manual curated ground truth as the test portion of the dataset, and all the paintings not included in the manual evaluation are part of the training or validation portions of the dataset. The division between training and validation portions is based on a random 80:20 split.

4.2 Resulting Dataset: Statistics and Analysis

In total, MELArt covers 1,616 artworks split in training (1,188), validation (328) and test (100) sets.

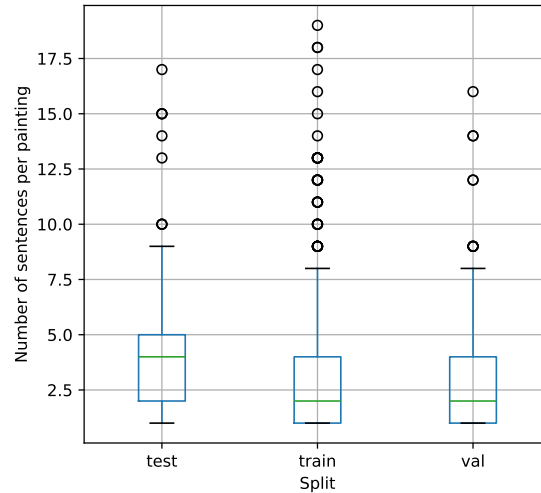
The dataset resulting from the process described above was constructed from 7,170 visual and 14,526 contextual sentences from Artpedia, contrasted with the matching of 2,118 visual and 2,716 contextual sentences in MELArt. The reduction in number of sentences is due to many sentences not directly mentioning the depicted entities (e.g., “The couple had been married for only a brief period”). The numbers also show a slightly stronger representation of visual sentences (44%) in MELArt as compared to Artpedia (33%). This is expected given that depicted entities are more likely to be mentioned in visual than contextual sentences.

The final MELArt statistics after introducing the manual annotations are:

- Number of mentions: 6,585
- The training, validation, and test splits correspond to: 4,632, 1,308, and 645 mentions respectively.
- 2,916 mentions are associated to visual sentences and 3,669 to contextual sentences.

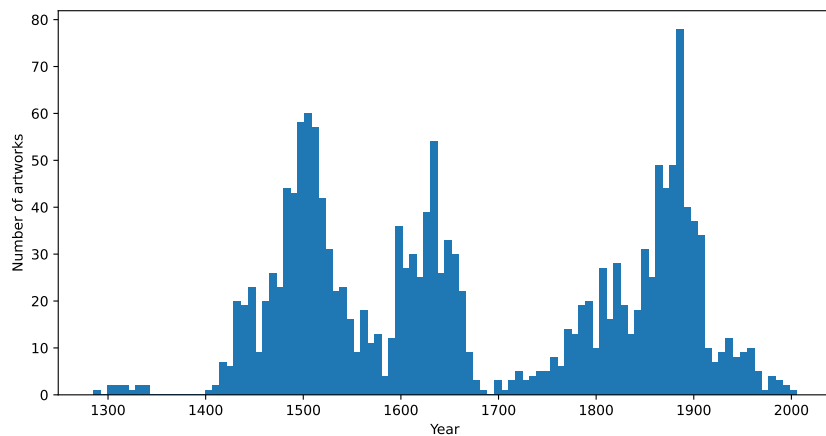
8:12 MELArt: A Multimodal Entity Linking Dataset for Art

In Figure 4, we can see that although there are especially long contextual sentences, the trend in terms of number of mentions and number of words per sentence remains similar among visual and contextual sentences.



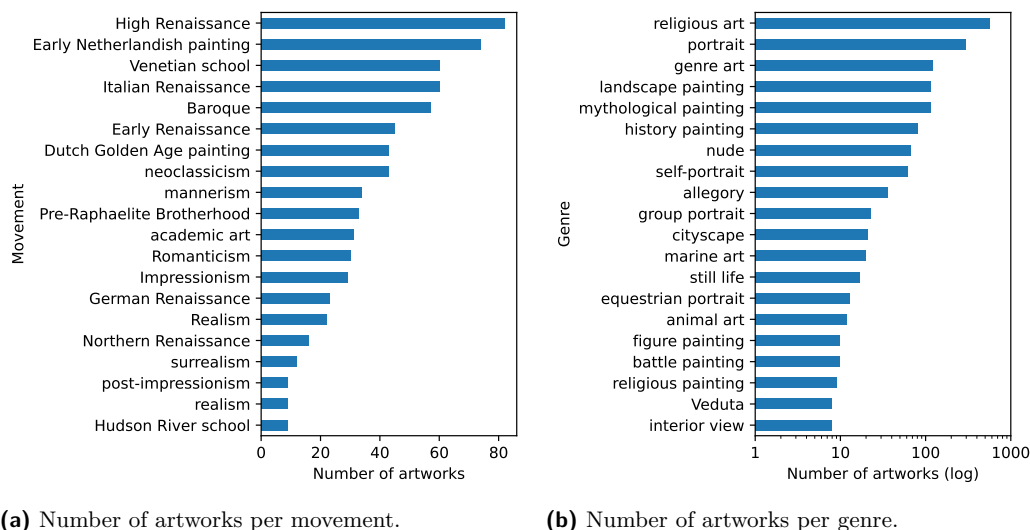
■ **Figure 5** Number of sentences per artwork by dataset split.

Considering the split of the dataset, we can see in Figure 5 that in terms of number of sentences for each artwork, the automatically annotated training and validation sets show similar trends. The manually annotated test set, however, has slightly more annotated sentences per artwork. This can be interpreted as a result of the lower recall of the automatic annotations, resulting in less mentions per artwork in the automatic annotations as compared to the manually annotated test set. When training a model with MELArt, this does not impose a problem in terms of the format of the training and validation data points, but raises an interesting challenge about linking potentially unseen surface forms that can be found in the test set.



■ **Figure 6** Artworks in MELArt by year of inception according to Wikidata.

In terms of the artworks in the dataset, we can leverage the information in Wikidata to understand potential biases in the dataset sources and generated annotations. Figure 6 shows a histogram representing the periods in which the included artworks were created⁵. We can see three strongly represented periods: one covering the 15th and 16th centuries, one for the 17th century, and lastly the 19th century.



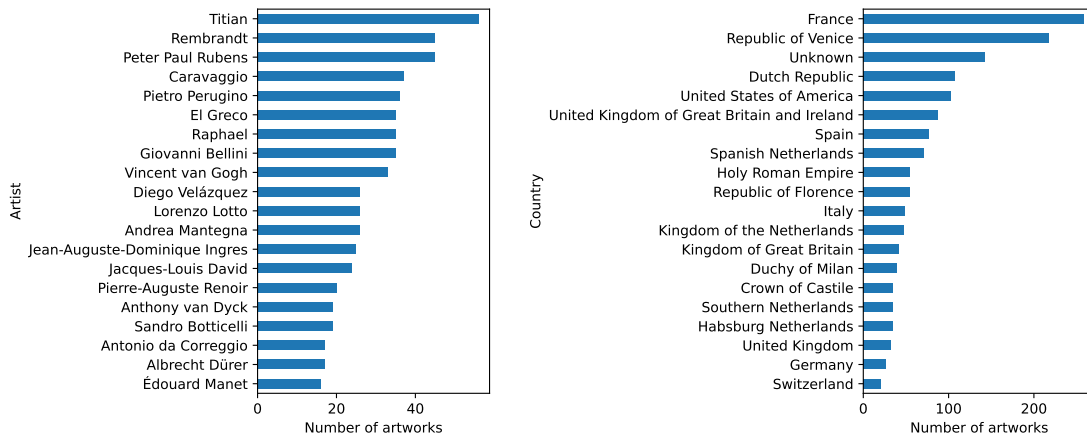
■ **Figure 7** Top 20 most represented movements and genres in MELArt.

By observing the movements and genres in Figure 7, it is clear that the *Renaissance* is the most common movement in the dataset, being also tightly related with the *religious art* genre. The genres show a stronger skewness towards *religious art* and *portrait*, which might be a result of focusing on named entities in MELArt. Analyzing the category distribution of Artpedia paintings before and after matching with the Wikidata “Depicts” predicate, reveals interesting distribution differences. In the Artpedia dataset, “religious art” (660 paintings), “portrait” (585), and “genre art” (259) are the most popular genres, while in the MELArt dataset, the dominance of “religious art” (571) and “portrait” (295) persists. Notably, “genre art” sees a significant drop from 259 to 121 in the matched dataset. Categories like “mythological painting” and “history painting” show a smaller decrease in representation. The matched dataset also exhibits a drastic reduction in categories like “landscape art” and “animal art”, highlighting a potential under-representation of these genres in MELArt due to the lack of named entities depicted in them. This comparison reveals a skew in MELArt’s representation of art, with a strong focus on religious and portrait art, while other genres like “landscape art” being less prominently featured due to the scarcity of named entities in them. This reflects possible biases and limited art genre diversity in both Artpedia and MELArt.

In Figure 8 we analyze the origin of the artworks according to the artists, where it is clear that the dataset is strongly skewed towards European art. Another clear bias in the dataset is related to the gender of the artists: out of 468, 413 are male, 27 are female and 1 is a non-binary/queer artist, according to gender information in Wikidata.

⁵ Making use of Wikidata’s inception date (P571).

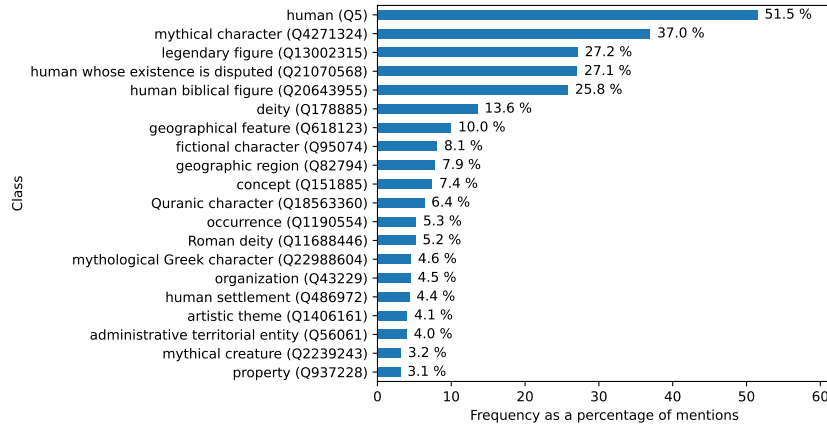
8:14 MELArt: A Multimodal Entity Linking Dataset for Art



(a) Number of artworks per artist.

(b) Number of artworks per artist's nationality.

■ **Figure 8** Top 20 most represented artists and their nationalities in MELArt.



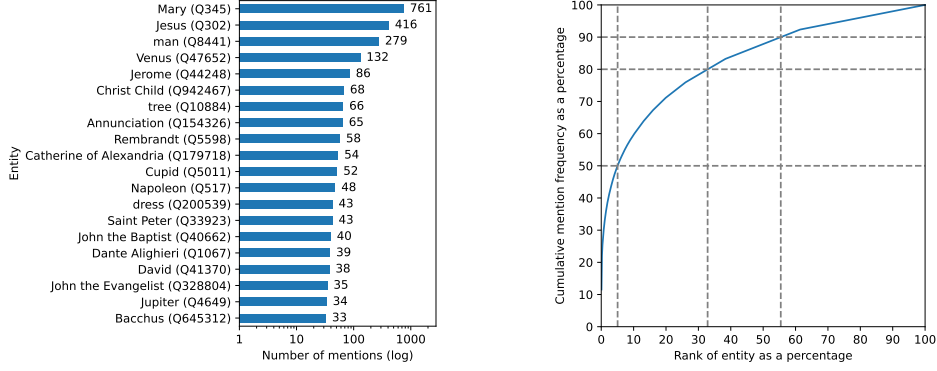
■ **Figure 9** Top 20 most represented classes in MELArt's mentions.

We also illustrate statistics about the depicted subjects in the annotations in Figures 9 and 10. We observe that although entities from different classes⁶ are mentioned in the sentences it is more common to find *humans* and its subclasses than others, e.g., *geographical features*.

In terms of concrete entities, aligned with the *genres* seen in Figure 7, religious entities are very prominent, with *Mary* (Q345) being the most frequently mentioned entity in the dataset. We also see that the dataset is skewed towards a group of entities, having half of the mentions linked to the most frequent five percent of the 1,306 mentioned entities, and eighty percent of the mentions being covered by one third of all the entities.

All the artworks have an associated image in the dataset. Roughly half of the full candidate set contain at least one image. In the subset of mentioned candidates more than 90% contain at least one image.

⁶ We restrict our analysis to Wikidata classes directly related to the mentions.



(a) Number of mentions by entity (Top 20).

(b) Skewness of the entity representation.

■ **Figure 10** Sparsity of the entity mentions.

4.3 Evaluation of Entity Linking Baselines using MELArt

In this section, we present the evaluation of tree entity linking methods using the MELArt dataset. To this end, each method ranks the candidates according to its specific scoring function, and we leverage on common ranking evaluation metrics to compare them. The resulting scores are sorted in descending order to compute Hits at k ($H@k$), Mean Reciprocal Rank (MRR), and Mean Rank (MR). These metrics are computed as follows:

$$H@k = \frac{1}{N} \sum_i I(\text{rank}(i) < k) \quad (1)$$

$$MRR = \frac{1}{N} \sum_i \frac{1}{\text{rank}(i)} \quad (2)$$

$$MR = \frac{1}{N} \sum_i \text{rank}(i) \quad (3)$$

The metric $H@k$ reflects the presence of the correct entity within the top k entities ranked by score. MRR denotes the average of the inverse of the rank of the correct entity, and MR represents the average rank of the correct entity relative to all entities. Thus, higher values of $H@k$ and MRR correspond to better performance, whereas for MR , a lower value denotes superior performance. The measured ranks are computed per mention, meaning that if a sentence has two entities mentioned, each one is ranked independently.

In order to showcase the use of our dataset to evaluate multimodal entity linking method effectiveness on art data, we tested one entity and relation linking tool for Wikidata and trained and tested two state-of-the-art (SOTA) entity-linking methods using our MELArt dataset.

MIMIC (KDD-2023)

In this work [6], the authors introduce a novel framework, called MIMIC, to improve multimodal entity linking which connects web content to a multimodal knowledge graph. Addressing the limitations of previous methods in handling abbreviated texts and implicit visual cues, MIMIC uses an input and feature encoding layers and three specialized interaction units: TGLU (Text-based Global-Local Interaction Unit) for textual context, VDLU (Vision-based Dual Interaction

Unit) for visual cues, and CMFU (Cross-modal Fusion-based Interaction Unit) for cross-modal fusion. This approach, validated on three benchmark datasets, significantly outperforms existing models, showcasing its efficiency in extracting and integrating complex multimodal data. In our experiments, we evaluate three configurations for MELArt using this model: “MIMIC” refers to including the painting and candidate images when present for the multi-modal entity linking tasks, “MIMIC (no cand. image)” excludes the candidate images and uses only the paintings, and “MIMIC (no images)” where MIMIC only uses the textual information. The textual information that describes the candidates is the concatenation of the main label, the entity description, and its types. For instance, “Mary” (Q345) is described as the string “Mary. mother of Jesus. Types: human biblical figure”.

For training the bi-encoder, the following hyperparameters were used:

- Learning rate: $1e^{-5}$
- CLIP model: openai/clip-vit-base-patch32
- Batch size: 128
- Maximum number of epochs: 20

BLINK (EMNLP-2020)

This work [16] introduces a new methodology for text-based entity linking, particularly targeting zero-shot scenarios. This approach is notable for its ability to link text to entities in a knowledge base without having seen these entities during the training phase. The authors propose a two stage model. The first stage (bi-encoder) uses dense vector representations for entity retrieval, a method that significantly enhances scalability and efficiency. The second stage (cross-encoder), uses the top-k most similar entities discovered by the bi-encoder, and re-ranks them by jointly encoding the sentence (context) and candidate text. This innovative approach offers a more effective solution for zero-shot entity linking challenges, where traditional methods often struggle due to the lack of prior training data on new entities. In our experiments, we train and evaluate the bi-encoder, and use the best bi-encoder candidates to train a cross-encoder. It is important to note that there is a risk of propagating errors from the bi-encoder to the cross-encoder, especially during training. Thus, by using the best bi-encoder for training the cross-encoder we explore the upper bound of BLINK’s performance using MELArt. We train the bi-encoder using *bert-large-uncased*, extracting the top 30 candidates, and train the cross-encoder with those candidates using *bert-base-uncased*. For reporting MR and MRR we assume a rank of 500 for entities that have not been found in the top 30 candidates of a mention.

For training the bi-encoder the following hyperparameters were used

- Learning rate: $3e^{-5}$
- BERT model: bert-large-uncased
- Batch size: 8
- Maximum number of epochs: 10

For training the cross-encoder:

- Learning rate: $2e^{-5}$
- BERT model: bert-base-uncased
- Batch size: 2
- Maximum number of epochs: 5

FALCON 2.0 (CIKM-2020)

This tool presented in [8], and available through an API⁷, allows users to submit portions of text, and it automatically recognizes and links entities to and relations in the Wikidata knowledge graph. The authors propose to jointly recognize and link entities and relations. The recognition phase is based on the concept of N-Gram tiling based on English morphological rules, and the linking phase ranks triples of recognized entities and relations based on the background knowledge from the knowledge graph. We acknowledge that FALCON does not necessarily solve the same task as MIMIC and BLINK, due to its more complex task of both recognition and linking. Additionally, FALCON does not use the candidate set of MELArt, but the whole set of Wikidata entities as candidates, making the ranking exercise harder. However, in our evaluation experiments, we want to include FALCON as an off-the-shelf alternative to specialized entity linking models.

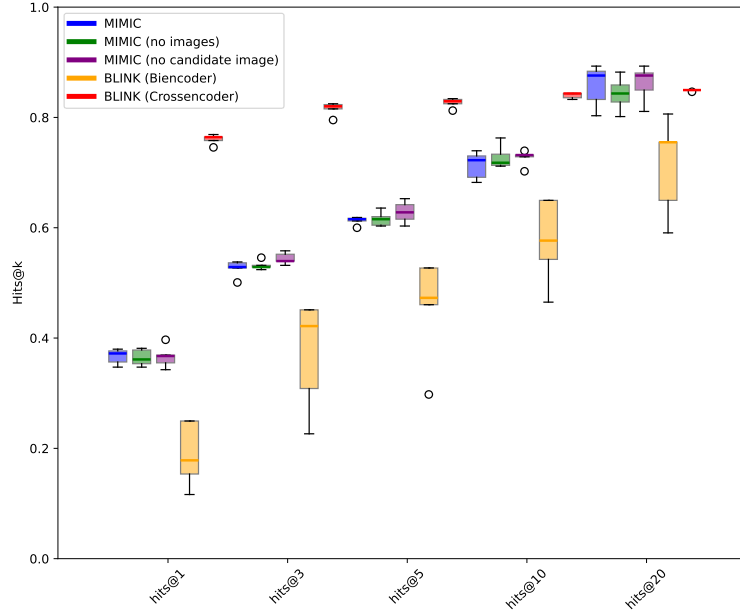
Entity Linking Results

■ **Table 2** Results of SOTA Entity Linking baselines on our dataset. The numbers enclosed by brackets correspond to the standard deviation for multiple runs.

Baselines	H@1	H@3	H@5	H@10	H@20	MR	MRR
BLINK (Bi-encoder)	0.19 (±0.06)	0.37 (±0.1)	0.46 (±0.09)	0.58 (±0.08)	0.71 (±0.09)	128.51 (±41.02)	0.32 (±0.07)
BLINK (Cross-encoder)	0.76 (±0.01)	0.82 (±0.01)	0.83 (±0.01)	0.84 (±0.01)	0.85 (±0.00)	76.25 (±0.1)	0.79 (±0.01)
MIMIC	0.37 (±0.01)	0.53 (±0.01)	0.61 (±0.01)	0.71 (±0.03)	0.86 (±0.04)	50.03 (±15.49)	0.48 (±0.01)
MIMIC (no cand. image)	0.37 (±0.02)	0.54 (±0.01)	0.63 (±0.02)	0.73 (±0.01)	0.86 (±0.03)	56.14 (±17.72)	0.49 (±0.01)
MIMIC (no images)	0.36 (±0.02)	0.53 (±0.01)	0.62 (±0.01)	0.73 (±0.02)	0.84 (±0.03)	71.81 (±24.68)	0.48 (±0.01)
FALCON 2.0	0.09	0.13	0.16	0.16	0.19	403.95	0.12

By executing these SOTA entity linking methods on our new dataset, we aim to assess the usefulness of MELArt for training these models, as well as understanding how challenging MELArt is as a test dataset. In our work, BLINK [16] and MIMIC [6] are chosen as the SOTA baselines. BLINK, a SOTA method for textual entity linking tasks, provides a rigorous standard against the text-processing capabilities of new systems to be evaluated. MIMIC, a SOTA multimodal entity linking, is employed alongside BLINK to test the integration and interpretation of both textual and visual features in the entity linking task. Additionally, FALCON 2.0 [8], serves as a text-based entity recognition and linking baseline. FALCON inference model is different from specialized entity-linking models like MIMIC and BLINK because instead of being provided with a mention intended to be linked, the input is only the raw text. FALCON recognizes the entities and relations in the texts and ranks candidates for each recognized span. Thus, in our evaluation we run FALCON with the test set of MELArt, and then for each mention in the evaluation set, we match the surface form with the mention text. If the surface form of the recognized entity overlaps with the text in MELArt’s mention, we evaluate the top k candidates of FALCON. We only consider a hit, if the mention can be matched to the surface form of an entity span from

⁷ <https://labs.tib.eu/falcon/falcon2/api-use>



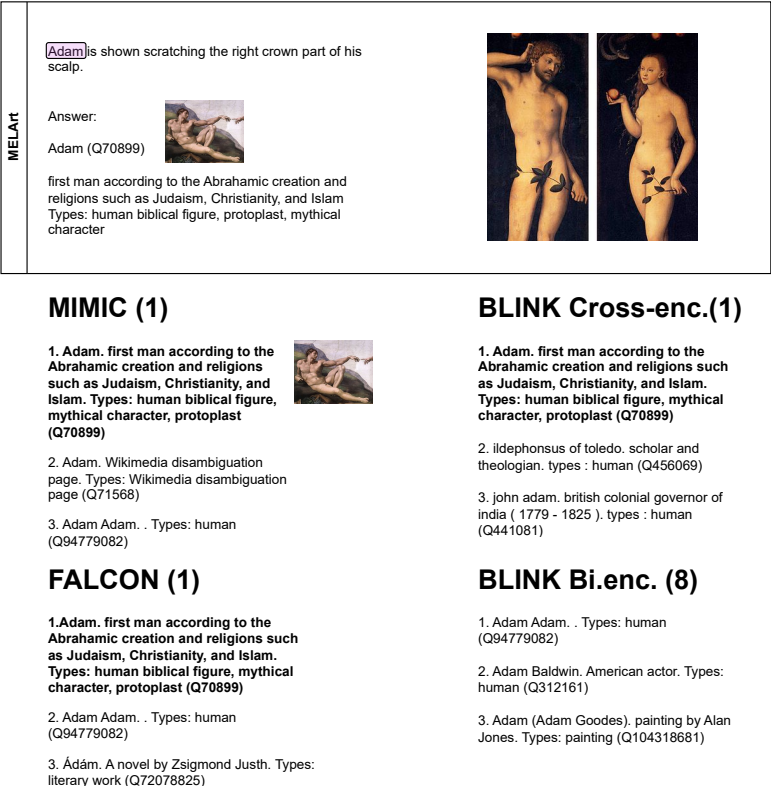
■ **Figure 11** Hits@K for different values of k in different runs.

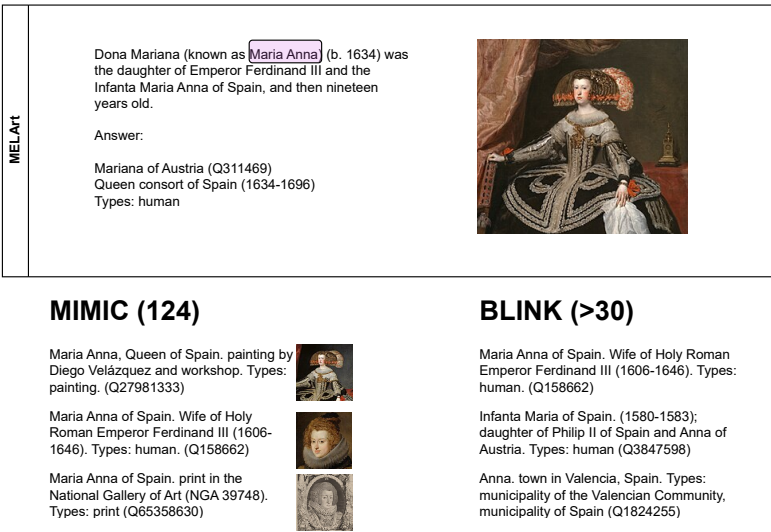
FALCON. Given that not all mentions can be matched between FALCON and MELArt, and that we only consider the first 50 ranked candidates from FALCON, similar to BLINK we assume a rank of 500 when the ground truth could not be found in the first 50 results.

Based on Table 2, MIMIC and BLINK exhibit distinct performance characteristics as observed across the various evaluation metrics considered. In all measures except H@20, BLINK (Cross-encoder) outperforms the other baselines, suggesting that textual information remains the most important feature to disambiguate between candidates. However, comparing BLINK and FALCON, we observe that training and fine-tuning the model, which is the setup for BLINK experiments, strongly improves the linking results. However, comparing BLINK and FALCON in absolute numbers is not a fair comparison, given the fact that FALCON also performs the recognition phase that can propagate errors to the linking phase. In fact, in our experiments, only 20% of the MELArt mentions could be matched to FALCON recognized spans, thus hindering FALCON’s measures in general. What is very clear, is that the cross-encoder training has a strong positive impact on the distinct measures.

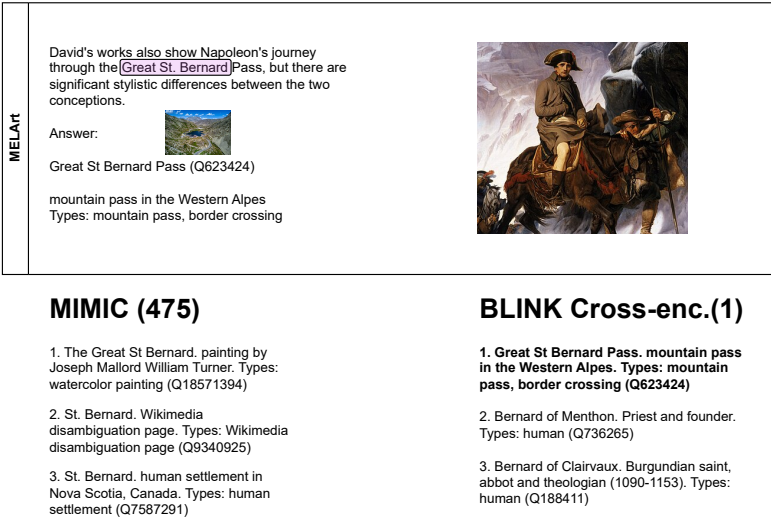
MIMIC is the only multi-modal baseline in our experiments, and from the three configurations that we used for MIMIC, we observe that including images does not significantly improve performance, suggesting that the difference in style between multiple representations of the same entity confuses the computer vision components of the linking model.

Finally, observing Figure 11, where we compare the variance in results with different seeds for each model, we note that the bi-encoder’s performance varies a lot depending on the initial parameter values. In fact, with one of the seeds that we tested, we obtained results close to zero for all measures, but we excluded that test from the compiled results. This variability of the bi-encoder is a high risk for the cross-encoder model because it relies on the bi-encoder predictions to perform the final re-ranking. Contrary to the bi-encoder, the BLINK cross-encoder has very low variety for a fixed set of candidates.





■ **Figure 13** Top-ranked entities for the mention *Maria Anna*. On top the ground truth as defined in the manual annotations of MELArt, on the bottom the predictions and the ground truth rank in parentheses.



■ **Figure 14** Top-ranked entities for the mention *Great St. Bernard*. On top the ground truth as defined in the manual annotations of MELArt, on the bottom the predictions and the ground truth rank in parentheses.


MELArt

Finally, at the top left is Simon of Cyrene, his face upside upturned.

Answer:
Simon of Cyrene (Q328739)

human biblical figure in Matthew 27:32, man who was forced by the Romans to carry the cross of Jesus. Types: human biblical figure





MIMIC (1)

1. Simon of Cyrene. human biblical figure in Matthew 27:32, man who was forced by the Romans to carry the cross of Jesus. Types: human biblical figure (Q328739)
2. Simon of Cyrene carries the cross. fifth Station of the Cross. Types: statio (Q114315626)
3. Order of Simon of Cyrene. . Types: award (Q7100546)

BLINK Cross-enc.(>30)

1. Saint Catherine of Alexandria. sculpture by David Zürn. Types: sculpture (Q124993845)
2. St Catherine of Alexandria. painting by Bernhard Strigel. Types: painting (Q64789016)
3. Simon Simon. Swiss topographer (1857-1925). Types: human (Q28082980)

■ **Figure 15** Top-ranked entities for the mention *Simon of Cyrene*. On top the ground truth as defined in the manual annotations of MELArt, on the bottom the predictions and the ground truth rank in parentheses.

consider the ground truth in the top-30 candidates. This is a typical issue of error propagation in models like BLINK in which a lighter model (Bi-encoder) first filters a set of candidates for a stronger model to re-rank in the second stage.

5 Conclusion

In this paper, we have introduced a novel multimodal entity linking dataset, characterized by its reliability and thorough evaluation tested over various entity linking baselines, both in a multimodal and in a text-based setting. The dataset is automatically extracted from multiple sources and supplemented with meticulous human annotations. The MELArt dataset, carefully curated and rigorously tested, presents a novel resource that challenges traditional NEL approaches and offers rich opportunities for advancing the field of entity linking in the art domain.

The results obtained from our experimental evaluation reveal that our dataset poses a significant challenge to existing general-domain pretrained entity linking models, and that training and fine-tuning models enhance their in-domain linking performance. This makes our dataset MELArt a valuable asset for researchers and practitioners in this domain. The results also indicate that there is room for improvement in the usage of images which are particularly important for the art domain, and that specialized models might gain better performance.

References

- 1 Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. Artemis: Affective language for visual art. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11569–11579. Computer Vision Foundation / IEEE, 2021. doi:10.1109/CVPR46437.2021.01140.
- 2 Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Multimodal entity linking for tweets. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*, volume 12035 of *Lecture Notes in Computer*

- Science*, pages 463–478. Springer, 2020. doi: 10.1007/978-3-030-45439-5_31.
- 3 Sofia Baroncini, Bruno Sartini, Marieke van Erp, Francesca Tomasi, and Aldo Gangemi. Is dc:subject enough? A landscape on iconography and iconology statements of knowledge graphs in the semantic web. *Journal of Documentation*, 79(7):115–136, March 2023. doi:10.1108/JD-09-2022-0207.
- 4 Hannah Bast and Björn Buchhold. Qlever: A query engine for efficient sparql+text search. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, CIKM '17, pages 647–656, New York, NY, USA, 2017. ACM. doi:10.1145/3132847.3132921.
- 5 Sabine Lang and Björn Ommer. Transforming information into knowledge: How computational methods reshape art history. *Digital Humanities Quarterly*, 15(3), 2021. URL: <http://www.digitalhumanities.org/dhq/vol/15/3/000560/000560.html>.
- 6 Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1583–1594. ACM, 2023. doi:10.1145/3580305.3599439.
- 7 Roberto Pirrone, Vincenzo Cannella, Orazio Gambino, Arianna Pipitone, and Giuseppe Russo. Wikiart: An ontology-based information retrieval system for arts. In *Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy, November 30-December 2, 2009*, pages 913–918. IEEE Computer Society, 2009. doi:10.1109/ISDA.2009.219.
- 8 Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. Falcon 2.0: An entity and relation linking tool over wikidata. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, CIKM '20, pages 3141–3148, New York, NY, USA, 2020. ACM. doi:10.1145/3340531.3412777.
- 9 Bruno Sartini. IICONGRAPH: improved iconographic and iconological statements in knowledge graphs. In *The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part II*, volume 14665 of *Lecture Notes in Computer Science*, pages 57–74, Cham, 2024. Springer. doi:10.1007/978-3-031-60635-9_4.
- 10 Alejandro Sierra-Múnera, Linh Le, Gianluca Demartini, and Ralf Krestel. MELArt Dataset. Dataset (visited on 2024-12-10). URL: <https://doi.org/10.48610/8a1ccdf>.
- 11 Alejandro Sierra-Múnera, Linh Le, Gianluca Demartini, and Ralf Krestel. MELArt Dataset Generation. Software, swhId: `swh:1:dir:ec4380448f4087c011040d0e3dca7832baa11182` (visited on 2024-12-10). URL: <https://github.com/HPI-Information-Systems/MELArt>, doi:10.4230/artifacts.22529.
- 12 Alejandro Sierra-Múnera, Linh Le, Gianluca Demartini, and Ralf Krestel. MELArt Experiments. Software, swhId: `swh:1:dir:203f2a69c5bc9064db3873a0160ca52f62095c25` (visited on 2024-12-10). URL: https://github.com/HPI-Information-Systems/MELArt_experiments, doi:10.4230/artifacts.22616.
- 13 Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing - ICIAP 2019 - 20th International Conference, Trento, Italy, September 9-13, 2019, Proceedings, Part II*, volume 11752 of *Lecture Notes in Computer Science*, pages 729–740. Springer, 2019. doi:10.1007/978-3-030-30645-8_66.
- 14 Peng Wang, Jiangheng Wu, and Xiaohang Chen. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 938–948. ACM, 2022. doi:10.1145/3477495.3531867.
- 15 Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4785–4797. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.acl-long.328.
- 16 Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6397–6407. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.519.
- 17 Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahim, Nanne van Noord, and Giorgos Tolias. The met dataset: Instance-level recognition for artworks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL: <http://cmp.felk.cvut.cz/met/>.