Multimedia Tools and Applications (2025) 84:39259–39282
https://doi.org/10.1007/s11042-024-20561-9

1250: HERITAGE PRESERVATION IN THE DIGITAL AGE: ADVANCES IN MACHINE LEARNING, MONOMODAL AND MULTIMODAL PROCESSING, AND HUMAN-MACHINE INTERACTION

EUFCC-340K: A faceted hierarchical dataset for metadata annotation in GLAM collections

Francesc Net[1] · Marc Folia[2] · Pep Casals[2] · Andrew D. Bagdanov[3] · Lluis Gómez[1,4]

Received: 16 April 2024 / Revised: 21 November 2024 / Accepted: 20 December 2024 /
Published online: 29 January 2025

Abstract
In this paper, we address the challenges of automatic metadata annotation in the domain of Galleries, Libraries, Archives, and Museums (GLAMs) by introducing a novel dataset, EUFCC-340K, collected from the Europeana portal. Comprising over 340,000 images, the EUFCC-340K dataset is organized across multiple facets – Materials, Object Types, Disciplines, and Subjects – following a hierarchical structure based on the Art & Architecture Thesaurus (AAT). We developed several baseline models, incorporating multiple heads on a ConvNeXT backbone for multi-label image tagging on these facets, and fine-tuning a CLIP model with our image-text pairs. Our experiments to evaluate model robustness and generalization capabilities in two different test scenarios demonstrate the dataset's utility in improving multi-label classification tools that have the potential to alleviate cataloging tasks in the cultural heritage sector. The EUFCC-340K dataset is publicly available at https://github.com/cesc47/EUFCC-340K.

Keywords Automatic metadata annotation · Hierarchical datasets · Image tagging · Cultural heritage · GLAM

✉ Lluis Gómez
Luis.Gomez@uab.cat

Francesc Net
fnet@cvc.uab.cat

Marc Folia
marc.folia@nubilum.es

Pep Casals
pep.casals@nubilum.es

Andrew D. Bagdanov
andrew.bagdanov@unifi.it

1 Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona 08290, Catalunya, Spain

2 Nubilum, Gran Via de les Corts Catalanes 575, 1r 1a, Barcelona 08011, Catalunya, Spain

3 Media Integration and Communication Center, University of Florence, Florence 50134, Italy

4 Department of Computer Science, Universitat Autonoma de Barcelona, Edifici Q, 08193 Bellaterra (Cerdanyola del Vallès), Spain

1 Introduction

GLAMs (Galleries, Libraries, Archives, and Museums) are institutions that compile and maintain collections of diverse cultural heritage assets in the public interest. Their main functions are to preserve assets with cultural and historical value and make them available to researchers and the broader public alike. However, the traditional framework for asset cataloging within GLAMs relies heavily on manual metadata annotation, a complex and labor-intensive process undertaken by expert professionals adhering to stringent standards. This method, while effective, presents challenges to scalability and efficiency.

The motivation behind our work lies in the recognition of these challenges and the need for innovative solutions. We aim to develop advanced tools that streamline the cataloging workflow and make a step forward in the use of artificial intelligence to help GLAM cataloging experts in the inventory process for cultural heritage collections.

To this end we take an image tagging perspective. Image tagging consists of automatically assigning a list of tags (or category labels) to a given cultural asset's image, referring to words that describe its visual content and/or cultural context. This is a multi-label classification task in which an image usually belongs to more than one class (i.e. class labels are not mutually exclusive).

However, using state-of-the-art neural networks for automatic metadata annotation in the GLAM domain comes with its own set of challenges. First, the number of categories we need to consider can be very large – for example, the Getty Research Institute's "Art & Architecture Thesaurus"[1] has more than 165, 000 terms – and the number of training examples for most of them can be very small (long tail) [1, 2]. On the other hand, the problem of incomplete annotations or missing labels is caused by the fact that two human annotators may consider different subsets of labels relevant for the same image even though both annotations may be equally valid [3].

To address these challenges, we have created a novel dataset specifically tailored for image tagging in the GLAM domain. Our dataset, named EUFCC-340K (Europeana Facets Creative Commons), was collected from the Europeana portal,[2] which aggregates a diverse range of GLAM European institutions, encompassing galleries, libraries, archives, and museums. EUFCC-340K comprises a total of 383, 695 images of a rich variety of cultural heritage assets, including artworks, sculptures, numismatics artifacts, textiles and costumes, and more, reflecting the diverse nature of collections within GLAM institutions. As illustrated in Fig. 1 each image in the dataset is accompanied by manual annotations provided by domain experts, capturing a wide spectrum of descriptive tags.

In addition to the rich visual content of the images, the annotations provided for each image follow a structured framework consisting of several key facets. These facets include Materials, Object Types, Disciplines, and Subjects, each representing distinct aspects of the cultural heritage assets depicted in the images. Importantly, the terms used for annotation within each facet adhere to a hierarchical organization provided by the "Art & Architecture Thesaurus" (AAT). The Art & Architecture Thesaurus (AAT) [4] is a multilingual thesaurus widely recognized for its broad coverage of terminology relevant to the GLAM domain, particularly for describing Object Types, Materials, and other domain-specific attributes. In the context of Europeana and the Linked Open Data community, AAT has long been one of the reference vocabularies for enriching cultural heritage metadata.[3] In this way, we leverage the

[1] https://www.getty.edu/research/tools/vocabularies/aat/

[2] http://www.europeana.eu

[3] Europeana enriches its data with the Art and Architecture Thesaurus

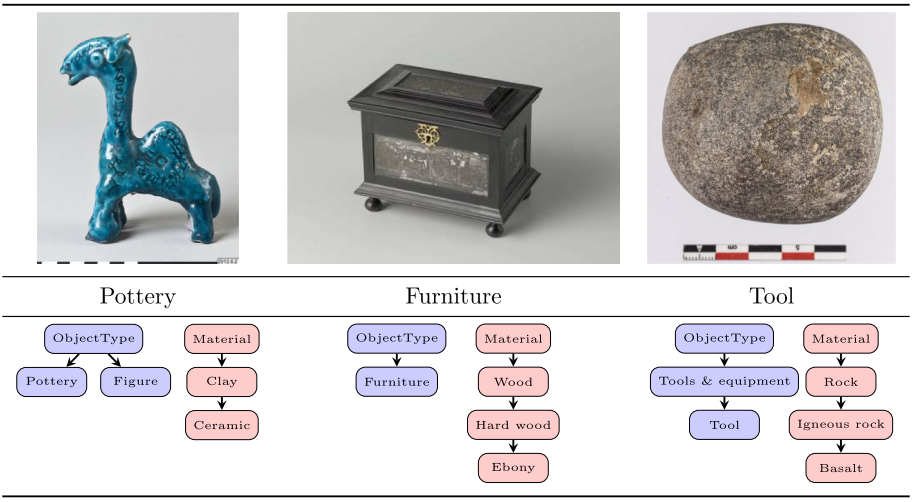


Fig. 1 Sample records from the EUFCC-340K dataset. Each image in the dataset is annotated across different facets of the Getty "Art & Architecture Thesaurus" (Some nodes are omitted for visualization purposes in this figure)

domain knowledge encapsulated within the thesaurus, thereby enhancing the interpretability and utility of our dataset.

The hierarchical structure of the dataset annotations provides inherent advantages in mitigating challenges associated with sparse training data for certain categories in the long-tail, particularly those situated at the leaf nodes of the hierarchy. Thanks to the hierarchical organization, the dataset facilitates more robust predictions for broader categories at higher levels of the hierarchy, even when training data for specific subcategories may be limited. This characteristic not only enhances the applicability of the dataset in real-world use cases but also contributes to the effectiveness of machine learning models in handling the long tail distribution of metadata annotations within the GLAM domain.

To evaluate the effectiveness of our dataset and to benchmark the performance of multi-label classification models in the GLAM domain, we also have developed a set of baselines tailored to address the multifaceted nature of this task. Our approach extends a state-of-the-art image classification architecture by augmenting them with multiple heads, each dedicated to classifying images based on a distinct facet within our annotation framework. Specifically, we append separate multi-label heads to the end of the model, with each head corresponding to one of the facets: Materials, Object Types, Disciplines, and Subjects. This design allows the model to simultaneously predict multiple labels across different facets, capturing the diverse visual and contextual characteristics of cultural heritage assets represented in our dataset. By evaluating these baselines, we aim to establish a basis for assessing the performance of image tagging algorithms in the GLAM domain and to provide insights into the challenges and opportunities inherent in automatic metadata annotation within this domain.

The rest of this paper is organized as follows. Section 2 provides an overview of related work, particularly focusing on research related to hierarchical image classification. In Section 3, we present a detailed description of the EUFCC-340K dataset, including its collection process, annotation framework, and characteristics. We discuss the design of the baseline models in Section 4, and also describe our architecture and implementation. In Section 5 we report on our experimental setup and results, evaluating the performance of the proposed

baselines on our dataset, and proposing an assistant tool for GLAM cataloguers. We conclude in Section 6 with a summary of our findings and discussion of possible future research.

2 Related work

TinyImages [5] was one of the earliest attempts to create a vast and varied hierarchical dataset for computer vision, containing 80 million 32x32 low-resolution images collected from the internet by querying every word in the WordNet [6] hierarchy. However, the dataset was limited by its low resolution and the noise inherent in its uncurated collection process. Following TinyImages, the ImageNet [7, 8] dataset was also organized according to the WordNet hierarchy, offering over 14 million manually verified and high-resolution images categorized into more than 20,000 synsets (WordNet nodes). The dataset provides image-level annotation of a binary label indicating the presence or absence of an object class in the image. A smaller dataset for benchmarking computer vision hierarchical classification models was proposed by Seo et al. [9], by restructuring the Fashion-MNIST dataset [10] into a simple three-level balanced hierarchy to classify fashion items.

The large-scale Open Images [11] dataset for object detection provides 15M bounding box annotations that are created according to a semantic hierarchy of classes. The evaluation protocol measures Average Precision (AP) for leaf classes in the hierarchy as normally, while the AP for non-leaf classes AP is computed involving all its ground-truth object instances and all instances of its subclasses.

Hierarchical data also emerges inherently in more specific domains, such as animal or plant classification. The iNaturalist dataset [12, 13] for large-scale species classification also offers a hierarchical organization, albeit with a simpler two-level hierarchy consisting of superclasses (“Insects”, “Fungi”, “Plants”, etc.) and finer classes. On the other hand, the Herbarium 2021 Half-Earth Challenge Dataset [14] for species recognition features a challenging real-world image classification task. The dataset includes more than 2.5 million images of vascular plant specimens representing 64,500 taxonomic labels. In addition to labels for species-level and below, labels at higher levels in the taxonomic hierarchy are also included (family and order).

Diverging from prior works, our EUFCC-340K dataset is specifically designed to meet the unique challenges of Galleries, Libraries, Archives, and Museums (GLAMs) in cataloging cultural heritage assets. The hierarchical structure of the “Art & Architecture Thesaurus” (AAT) offers a domain-specific framework far beyond the capabilities of general-purpose datasets like ImageNet. Moreover, its multi-faceted hierarchy – spanning Materials, Object Types, Disciplines, and Subjects – goes beyond object recognition and includes annotations for materials from which objects are made. This not only involves recognizing the object itself but also understanding the materials, which presents a considerably more complex challenge.

3 The EUFCC-340K dataset

The EUFCC-340K dataset was compiled using the REST API provided by the Europeana portal,[4] a central repository for cultural heritage collections across Europe. Europeana aggregates a diverse array of cultural artifacts, multimedia content, and traditional records from numerous European institutions, offering a rich bed of resources for academic and research purposes. The platform’s comprehensive metadata for each item, detailing various attributes

[4] https://pro.europeana.eu/page/apis

and historical contexts, provides a unique opportunity for in-depth analysis and application in computer vision.

The hierarchical labeling structure of the EUFCC-340K dataset is derived from and aligned with the Getty "Art & Architecture Thesaurus (AAT)" with the help of domain experts. The Getty AAT is a continually evolving vocabulary, designed to support the cataloging, research, and documentation of art, architecture, and other cultural material. Its structured hierarchy enables the detailed classification of concepts and objects and reflects the complex relationships and attributes inherent in the cultural heritage domain. Figure 2 illustrates the description and hierarchical position of the AAT term "Numismatics" according to the AAT Online search tool.

Keyword searches and filtering Initial data collection involved performing keyword searches for the broad "Collections" categories, such as "Furniture", "Tool", "Costume", among others. Results were filtered to include only entries with available thumbnails and tagged with "Reusability: Open," ensuring that the dataset comprises images suitable for open research and application. We intentionally excluded categories such as "Photography" and "Painting" from the primary dataset to avoid the complexities introduced by the overlap of content and form descriptions in artistic representations. Such items were considered separately to maintain the dataset's focus.

Mapping Europeana concepts to Getty AAT In the context of the Europeana portal, the vocabulary terms used in metadata descriptions are represented with URIs referencing exact terms in other public controlled vocabularies such as AAT, Wikidata,[5] or the Library of Congress Subject Headings.[6] For each record retrieved, we extracted all the "concepts" with URIs referencing terms from the AAT thesaurus. We also collected the concepts automatically assigned by Europeana, provided they already had a declared or explicit alignment with AAT terms. For instance, the Europeana concept http://api.europeana.eu/entity/concept/46.json (Musical Instrument) explicitly aligns with the AAT term http://vocab.getty.edu/aat/300041620, as indicated by the "exactMatch" field in the metadata shown below:

Listing 1 Excerpt from Europeana Concept "Musical Instrument" JSON data. This example illustrates how Europeana concepts are explicitly aligned with AAT terms using the "exactMatch" field.

```
{
  "id": "http://data.europeana.eu/concept/46",
  "type": "Concept",
  "prefLabel": {
    "en": "Musical instrument"
  },
  "exactMatch": [
    "http://vocab.getty.edu/aat/300041620",
    "http://www.wikidata.org/entity/Q34379",
    "http://id.loc.gov/ontologies/bibframe/MusicInstrument
       "
  ],
  "note": {
    "en": "Device created or adapted to make musical
       sounds"
  }
}
```

[5] https://www.wikidata.org

[6] https://id.loc.gov/authorities/subjects.html

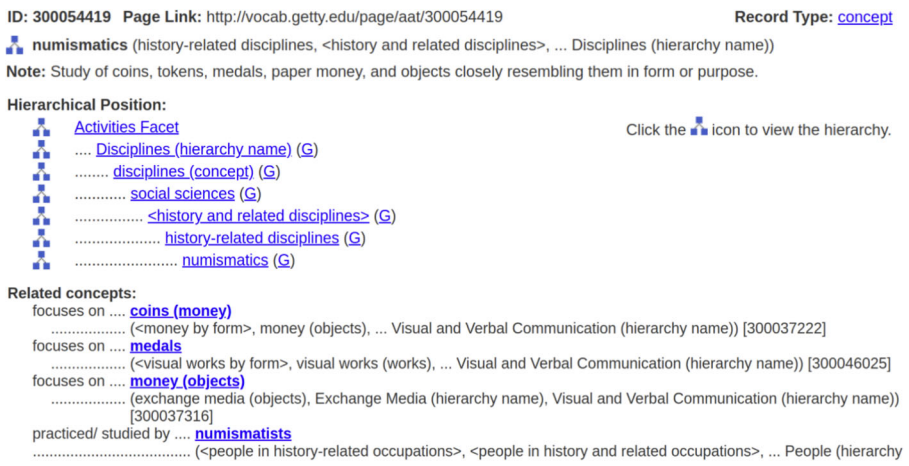


Fig. 2 Description and hierarchical position of the AAT term "Numismatics" according to the AAT Online search tool

All collected concepts were assigned to one of our four attributes—Materials, Object Types, Disciplines, and Subjects—based on their hierarchical inheritance in the Getty AAT thesaurus. Specifically:

- Concepts under the "Materials Facet" were assigned to the Materials attribute.
- Concepts under the "Objects Facet" were assigned to the Object Types attribute.
- Concepts under the branch `<visual works by subject type>` were assigned to the Subjects attribute.
- Concepts under the "Disciplines (hierarchy name)" were assigned to the Disciplines attribute.

Terms that did not belong to any of these branches were excluded from the dataset. This mapping process relied heavily on the hierarchical structure of Getty AAT, which provides clear semantic relationships for organizing terms. To simplify the structure, guide terms (enclosed in angled brackets) have been omitted from the hierarchy. In the AAT, guide terms are non-indexing descriptors used solely to group narrower terms for organizational purposes. They function as structural nodes that ensure logical grouping within the hierarchy, such as <styles, periods, and cultures by region>, but do not represent standalone concepts.

Linking all records to the same hierarchical structure following uniform rules provided a framework for organizing the dataset's labels, which facilitated the representation of complex relationships and attributes inherent in cultural heritage objects, and was critical for translating Europeana's diverse and rich metadata into a structured format aligned with Getty AAT's vocabulary.

Manual curation and quality assurance Despite the automated processes involved in data collection and mapping, manual curation plays an important role in ensuring the dataset's quality. The collected data was uploaded into a content management system with filtering and visualization capabilities, which facilitated a systematic review by domain experts. These experts checked for minimal coherence in concept usage across different data providers and identified terms that did not accurately describe the object represented in the image but rather referred to contextual features related to the object's provenance. Ambiguously used concepts were flagged for review, and in some cases, removed. Additionally, subsets of

the dataset containing inconsistencies or significant noise were removed to improve overall dataset reliability.

This manual intervention, though challenging, was crucial for maintaining the dataset's integrity and relevance to the GLAM domain. While the curation process was not exhaustive at the individual record level, the corrections and refinements significantly reduced noise and ensured a higher degree of consistency in the dataset's metadata. Still, we acknowledge that noisy annotations might still exist in specific categories.

Data provider information Each record was also annotated with information about the data provider, which, while not always directly corresponding to a single museum, offers insights into the source institution.

In total, our dataset consists of 346, 324 annotated images gathered from a total of 358 data providers. The contribution from these providers varies significantly, with some contributing as few as a single item, while others offer extensive inventories, including up to 40, 000 items. Figure 3, shows the items' distribution for the 30 most frequent data providers within the dataset.

Figure 4 illustrates the number of tags (class labels) per image and the overall frequency distribution of tags. Based on insights from this analysis, we introduced an additional filter to ensure a minimum of 10 images per tag. Furthermore, we excluded a specific museum, "Mobilier National," from the dataset due to an excessive number of tags per image, some of which were inaccurately labeled.

The number of unique tags for each facet is 33, 910, 8, and 276 for the "Classifications", "Object Types", "Subjects", and "Materials" respectively. It is worth noting the dataset items' are sometimes partially annotated, *i.e.*, not all images contain labels for all facets. Concretely, we have 78, 948 image annotations with one facet, 215, 196 with two facets, 51, 734 with three facets, and 446 with four facets. Moreover, in some instances, tag labels are assigned only up to a specific node in the hierarchy without necessarily reaching a leaf node. This reflects a practical and realistic use case, where not every image needs to be labeled down to the finest level of detail. Figures 1 and 5 illustrate several examples of the EUFCC-340K dataset.

To further illustrate the complex structure of our dataset, Fig. 6 provides a sunburst chart diagram visualizing the "Materials" facet hierarchy. This representation highlights the semantic richness and the multi-level nature of some hierarchies within our dataset. The diagram clearly shows the relationship between different levels of the hierarchy, from the

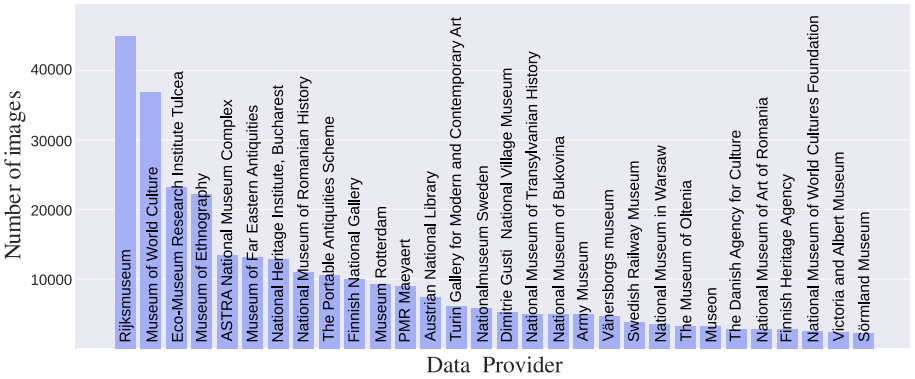
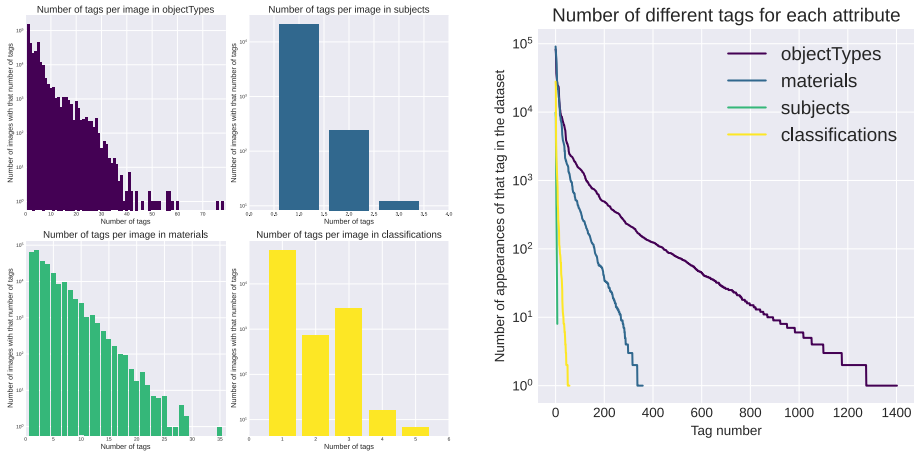


Fig. 3 Items' distribution for the 30 most frequent data providers within the EUFCC-340K dataset



(a) Number of tags per image (b) Number of instances of each tag

Fig. 4 EUFCC-340K statistics: tag count and tag length variation per image

central root node outwards to the child nodes, underscoring how data is distributed across various levels. This visualization not only aids in understanding the hierarchical relationships but also showcases the depth and breadth of possible annotations within the dataset.

Dataset splits Within the dataset, we have designated two distinct subsets to serve as the test data, collectively representing approximately 10% of the EUFCC-340K dataset. One subset is identified as the *Outter* test set, and the other as the *Inner* test set. The Outer test set aims to challenge the model with data that may differ significantly from the majority of the training dataset. To measure the divergence between image subsets' distributions, we utilized the Frechet Inception Distance (FID) [15], a widely used metric for assessing the similarity between two distributions of images. FID calculates this similarity by comparing the mean and covariance of features extracted from a pre-trained Inception v3 CNN model [16]. The features correspond to high-level representations of the images, capturing semantic content rather than pixel-level differences. The FID score is computed as:

$$d_F(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma'))^2 = \|\mu - \mu'\|_2^2 + \mathrm{tr}\left(\Sigma + \Sigma' - 2\left(\Sigma\Sigma'\right)^{\frac{1}{2}}\right) \quad (1)$$

where μ and Σ are the mean and covariance of the feature embeddings from one image distribution, and μ' and Σ' are those from another. Intuitively, FID quantifies how close the distributions of features are in a high-dimensional space. Lower FID scores indicate greater similarity between the two distributions, while higher scores suggest greater dissimilarity.

In our work, FID serves as a tool for identifying subsets of data that deviate from the overall dataset distribution. This allows us to detect potential subsets that challenge the model by exposing it to data distributions that differ from the training data. The metric's sensitivity to both mean shifts and covariance differences makes it especially effective for this purpose.

Therefore, we have carefully selected two data providers that account for 2% to 5% of the total data for the outer set. The first data provider in the Outer Test Set is chosen based on maximizing the FID with the entire dataset, ensuring that the image types are diverse and not exclusively from a single category (e.g., stamps). The second data provider in the Outer

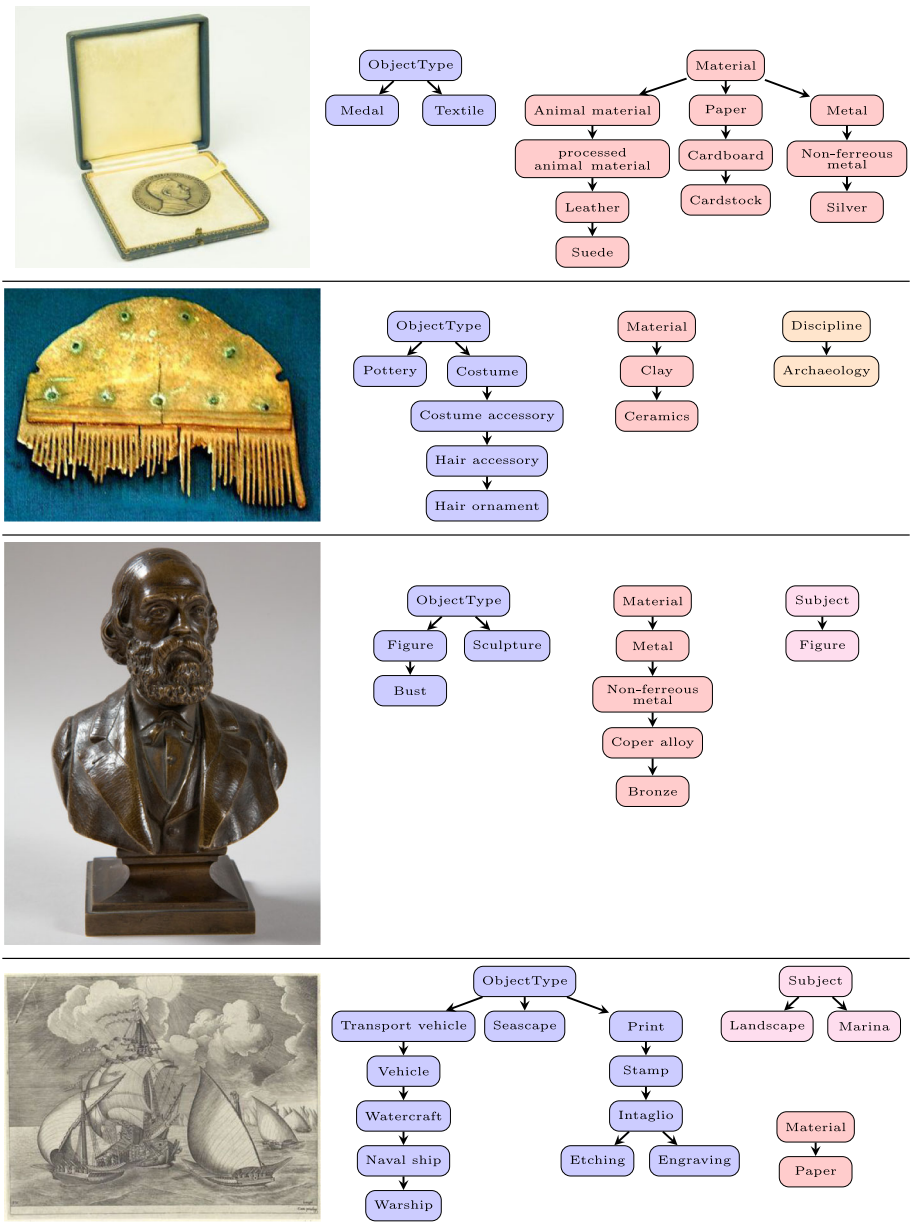


Fig. 5 Sample records from the EUFCC-340K dataset. Each image in the dataset is annotated across different facets' hierarchies of the Getty "Art & Architecture Thesaurus". Some nodes were omitted for visualization purposes

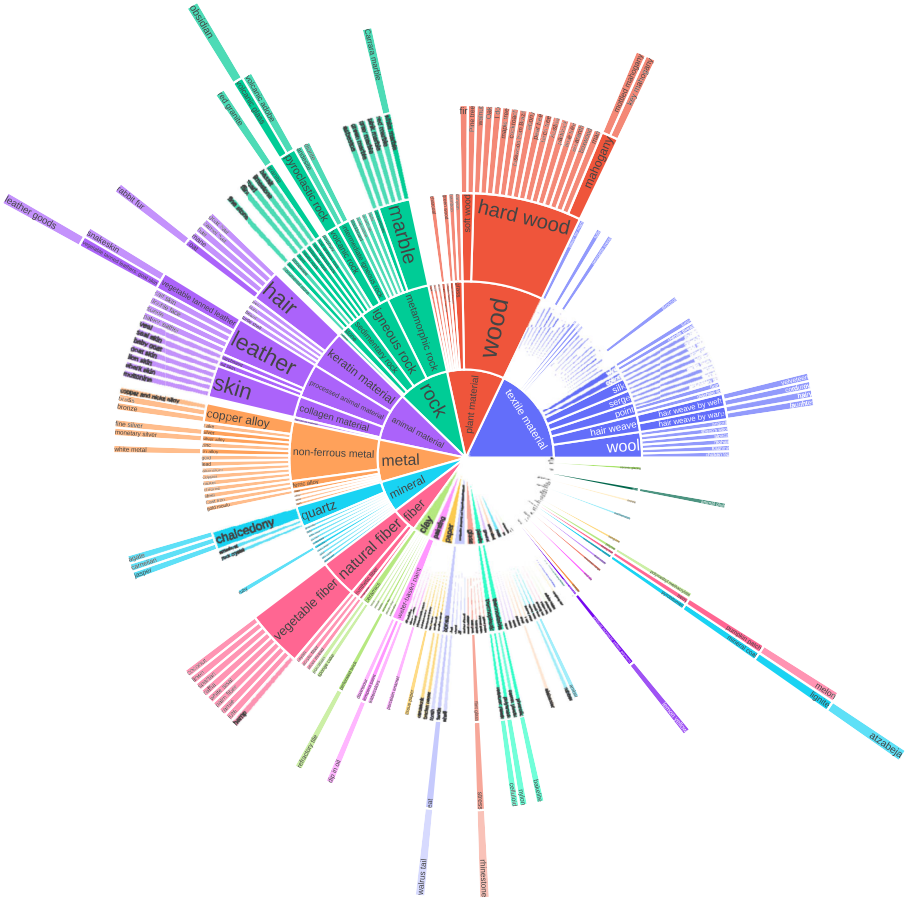


Fig. 6 Sunburst chart diagram visualization of the "Materials" facet hierarchy. The circle in the centre represents the root node, with the hierarchy expanding outwards from the center. It shows how the outer rings (child nodes) relate to the inner rings (parent nodes in the hierarchy), and how data is distributed across nodes

test set was selected as the one with FID to the first. The data providers conforming to the Outer test set are: "The Portable Antiquities Scheme" and "PMR Maeyaert".

For the Inner Test Set, we adhered to a balanced selection criterion, aiming for a 33% allocation of images for testing, with minimum thresholds set for each tag category (at least 5 images for classifications, object types, and materials, and 20 for subjects). This strategy ensures a diverse and representative sample of the dataset's breadth. Figure 7 illustrates the split of the training and the Inner Test Set, highlighting the methodological constraints applied to achieve balance and diversity. Although our dataset features tags with varying frequencies, our split strategy occasionally results in minor data distribution peaks, as depicted in the figure. The validation set, constituting another 10% of the dataset, is derived using similar criteria, with the remainder allocated for training. The whole dataset and official splits are made publicly available.[7]

[7] https://github.com/cesc47/EUFCC-340K.

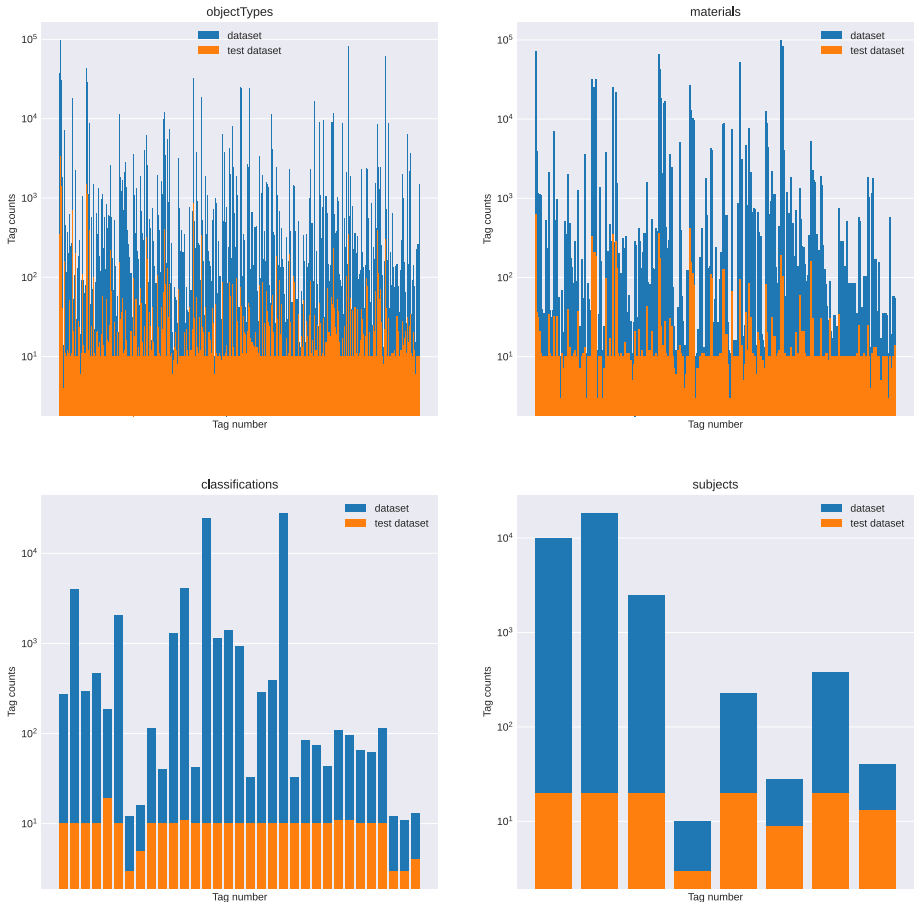


Fig. 7 EUFC-340K dataset split for the training set and Inner test set

4 Baseline methods

In this section, we introduce the baseline models that we use to benchmark performance on the proposed EUFCC-340K dataset for image tagging in the GLAM domain. We employ two distinct types of baseline models: vision-only models that leverage a state-of-the-art CNN architecture for image classification, and a state-of-the-art multi-modal model that projects visual and textual modalities into a shared common feature space. For each model, we detail their architecture and training strategies in the following subsections.

4.1 CNN-based vision models

Our vision-only baselines extend a state-of-the-art image classification CNN backbone, the ConvNeXT$_{base}$ model [17], by incorporating multiple classification heads. Each head is dedicated to one of the EUFCC-340K facets: Materials, Object Types, Disciplines, and Subjects. This multi-head setup enables the model to perform multi-label classification across various facets simultaneously. During training, we compute a separate loss for each classification

head, which allows specialized tuning of the model's ability to distinguish between class labels within each facet. This section details the various approaches and loss calculation methods used for each baseline within the vision-only architecture.

4.1.1 Multi-label classification

In this baseline, each output neuron in each of the heads has an individual sigmoid activation function ($\sigma(x) = \frac{1}{1+e^{-x}}$) that can be independently activated, allowing the model to predict the presence or absence of each tag individually. In other words, each output is a binary classifier that predicts whether each tag is activated (present) or not for the input image. Figure 8 illustrates the outputs of a multi-label head with 19 outputs, each output corresponds to a node in a given tag hierarchy with three levels, such as the node labeled as "0" is the parent of nodes {"00", "01", "02"}, and so on. The network is trained with Binary Cross Entropy (BCE) loss: $l(y, \hat{y}) = y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})$.

4.1.2 Softmax classification

In this baseline, we follow prior works [18, 19] that suggest that softmax classification can be very effective in multi-label settings with large numbers of classes. Softmax classification heads, as illustrated in Fig. 9, are the default choice for multi-class classification problems (where a given input belongs only to one single class), as the softmax output can be interpreted as a probability distribution over classes. In our context, the use of a softmax head in a multi-label setting involves the random selection of one tag from each attribute during the training phase and applying the cross-entropy loss function: $H(Y, \hat{Y}) = -\sum_{i=1}^{N} Y_i \log(\hat{Y}_i)$. At inference time, the predicted tags are sorted in descending order of probability.

4.1.3 Hierarchical softmax

The baseline models described thus far do not account for the hierarchy among class labels. They simplify (flatten) the class hierarchy, treating each output neuron either as an independent classifier in a multi-label setup or as part of a multi-class model that predicts a probability distribution over classes using a softmax output. In contrast, the Hierarchical Softmax baseline incorporates the class hierarchy into the labeling process. This approach provides context and structure, enabling the model to label images with multiple levels of granularity. We follow prior works [20] with the difference that we have several tagging heads instead of one. To perform classification with one of the hierarchical facets of the EUFCC-340K dataset we predict conditional probabilities at every node for the probability of each child node of

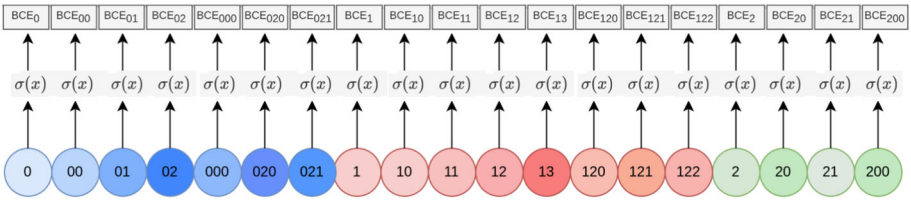


Fig. 8 An illustrative example depicting a multi-label classification head. Each output neuron behaves as a binary classifier for one label within a given hierarchy where the node labeled as "0" is the parent of nodes {"00", "01", "02"}, and so on

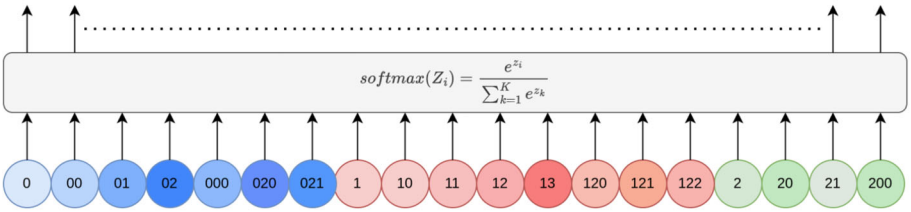


Fig. 9 Softmax classification head. The softmax output can be interpreted as a probability distribution over classes

a parent node given that parent node. As illustrated in Fig. 10 training this model involves computing a cross-entropy loss at every node that is relevant for a given input. For example, if a given input has the term "122" in its ground truth annotations the losses CE_{Root}, CE_1, and CE_{12} must be calculated and backpropagated, as depicted in Fig. 11.

Our hierarchical softmax models make use of weighted cross-entropy losses (assigning higher weights to minority classes and lower weights to majority classes) to facilitate unbiased training. Furthermore, in the experimental section, we also evaluate a modified version of the hierarchical softmax loss, where the loss is calculated per level instead of at individual tree nodes. This approach is adopted to prevent an emphasis on predictions that consistently follow a single-label path and to encourage a more diverse set of predictions.

4.2 Multi-modal baselines

Our multi-modal baselines involve fine-tuning a pre-trained CLIP model [21] using image-text pairs. The CLIP (Contrastive Language-Image Pre-training) model is designed to understand and generate representations that link textual and visual information, making it ideal for tasks that involve correlating images with descriptive texts. In our context, the text modality comprises all labels of a given image for a given facet, posing the challenge of effectively integrating them into a single text prompt. To address this, we explore diverse strategies for constructing text prompts, including: (1) using all raw tags separated by commas, (2) using a randomly selected single tag per prompt (similar to our approach in the softmax baseline), and (3) experimenting with template prompts tailored to distinct attributes by integrating

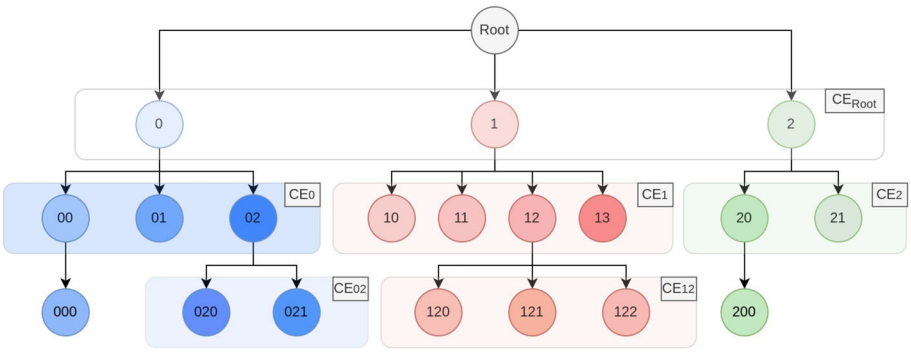


Fig. 10 An illustrative tree representation featuring a hierarchical softmax head, highlighting the specific nodes where cross-entropy (CE) losses might be computed

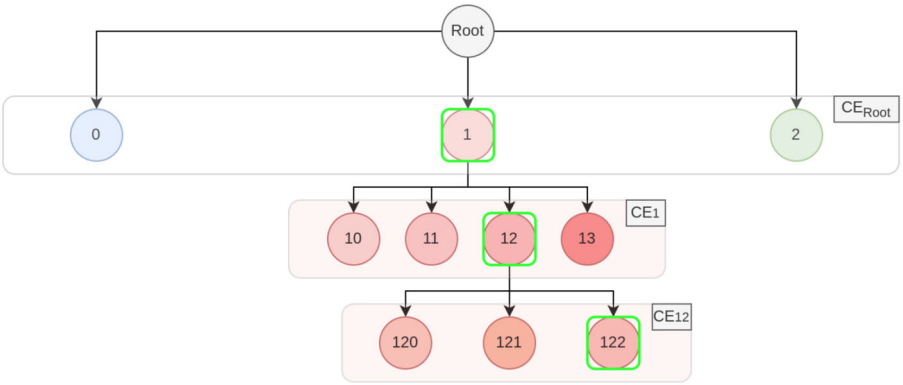


Fig. 11 An illustrative tree diagram demonstrating the computation of the loss function of a given input that is annotated with the term "122" in its ground-truth

relevant tags for each image. For enhanced clarity, consider the "Materials" facet, where the ground truth annotations for a given image might be "*metal*", "*non-ferrous metal*", and "*bronze*". The paired text for this image, depending on the chosen strategy, could be: (1) "`metal, non-ferrous metal, bronze`", (2) one of "`metal`", "`non-ferrous metal`", or "`bronze`", or (3) "`an image of an object made of <TAG>`" where `<TAG>` is a placeholder that can be substituted by any of the tags in the ground-truth. In the next section, we will compare the results of several CLIP baselines fine-tuned using all these prompt engineering strategies.

4.3 Ensemble classifier

Ensemble methods are learning algorithms designed to improve predictive performance by combining the outputs of multiple models. Instead of relying on a single model, ensembles leverage the diversity among multiple models to reduce errors and enhance robustness [22]. Ensembles often outperform individual models because the various errors of the ensemble constituents "average out," leading to improved generalization. This process is particularly useful for reducing the impact of overfitting by any single model.

Ensemble averaging is one of the most common techniques used in this context. It involves creating multiple models and averaging their outputs (e.g., class probabilities) to produce the final prediction. For an ensemble of E models and C classes, the predicted probability $\hat{p}_c$ for class c is calculated as: $\hat{p}_c = \frac{1}{E}\sum_{i=1}^{E} p_c^{(i)}$, where $p_c^{(i)}$ is the predicted probability for class c from the i-th model in the ensemble.

In our experiments, we employed an ensemble strategy that integrates outputs from multiple models with complementary strengths. Specifically, we combine predictions from: the multi-label classification model (using sigmoid activations), the softmax-based classification model – which provides a probability distribution across mutually exclusive classes, and (optionally) the CLIP model, which aligns image-text embeddings and produces a probability distribution over the class set.

During inference, the ensemble integrates these outputs by normalizing the sigmoid probabilities to ensure compatibility with the sotmax-based model. The overall prediction is then computed by aggregating the contributions from these models. When the CLIP model is included, its softmax probability distribution is also integrated into the ensemble.

5 Experiments

In this section, we first present the proposed evaluation framework for the EUFCC-340K dataset, then we give the implementation details for training all baseline models discussed in Section 4. Finally, we present and discuss the results obtained with all evaluated models, and present a tool designed to assist human expert annotators.

5.1 Evaluation metrics

Image tagging models are evaluated based on their ability to predict, for a given image, the relevance of each tag in the vocabulary, akin to creating a list of tags sorted by their predicted relevance. Then the model performance is measured using information retrieval metrics such as mean Average Precision (mAP). While mAP is widely used due to its ability to capture both precision and recall, it becomes less effective in datasets like EUFCC-430K where there is a high variability in the number of relevant tags across different images (see Fig. 4). Such variability can lead to inconsistent evaluations, as mAP may disproportionately favor queries with fewer relevant items or penalize those with more, due to its integrated recall component. To provide a more stable and meaningful evaluation, we have selected alternative metrics better suited to our dataset. We use R-Precision, Accuracy@K, and Average Rank Position, which focus on the precise assessment of how accurately tags are assigned without being affected by the variable number of tags per image. These metrics provide clear insights into both the precision of the tag assignments and the overall ranking quality of the model's outputs.

R-Precision assesses image tagging effectiveness by considering the precision of the top R tags, where R is the number of relevant tags known for a specific image. Formally:

$$\text{R-Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{r_i}{R_i}, \tag{2}$$

where N is the total number of images in the test set, R_i is the number of relevant tags for the i-th image according to the ground-truth, and r_i is the number of correct tags among the top R_i tags returned by the system. Note that this measure is equivalent to both precision and recall at the R_i^{th} position. In practical terms, if there are R tags relevant to an image, R-Precision looks at the top R tags returned by the system, counts how many of them are relevant (r), and calculates the ratio.

Accuracy@k assesses the performance of the model by measuring how often at least one of the ground- truth tags appears in the k highest-ranked predictions:

$$\text{Acc@k} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[\text{GT}_i \cap \text{top-K}_i], \tag{3}$$

where $\mathbb{1}[\text{GT}_i \cap \text{top-K}_i]$ takes the value 1 or 0 depending if at least one of the ground-truth tags (GT) is present in the top-K model predictions or not. In our experiments, we evaluate accuracy at $k = 1$ and $k = 10$.

Average rank position determines the average position of the ground truth labels in the output list, providing a quantitative measure of how well the model's predictions align with the true ranking. A lower AvgRankPos value indicates better performance. Formally,

AvgRankPos is defined as follows:

$$\text{AvgRankPos} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{R_i}\frac{Pos_{i,j}}{R_i}, \tag{4}$$

where R_i is the number of relevant tags for the i-th image according to the ground-truth, and $Pos_{i,j}$ is the ranking position of j-th tag for the i-th image.

5.2 Implementation details

Vision models All baselines were trained using the default pre-trained weights (from ImageNet) with four NVIDIA A40 GPUs, with a batch size of 128, and optimized using the AdamW optimizer, weight decay was set to 0.05, and a learning rate of 1e-4 during 25 epochs. The learning rate was scheduled using the cosine annealing strategy. The training process incorporated simple data augmentations: a random resized crop of 224 and a random horizontal flip. In addition, repeated augmentation sampling[8] (RA-Sampler) was used. Each training lasted 12 hours approximately.

Multi-modal models All baselines were trained with four NVIDIA A40 GPUs, with a batch size of 64, during 100 epochs, and utilizing a learning rate of 3×10^{-5}. Beta1 and beta2 parameters of Adam optimizer were configured at 0.9 and 0.99, respectively, accompanied by an epsilon value of 10^{-8} and a weight decay of 0.1. Other specifications included a warm-up period of 500 steps, gradient clipping with a norm of 1, and the adoption of the ViT-B-16 model architecture pre-trained on the WIT dataset [21]. The duration of fine-tuning varied based on the model's design-ranging from 3 to 7 days approximately-depending on whether it divided tags (resulting in a slower training process with more instances) or worked with all tags from an image collectively.

5.3 Automatic metadata annotation results

Tables 1 and 2 present the results of all baseline models for the Inner and Outer test sets respectively. We provide results for all baselines defined in Section 4 – Multi-label, Softmax, Hierarchical Softmax (H-Softmax), and CLIP models with different prompting strategies (all tags, one tag, and one tag with template prompt) – using all metrics described in Section 5.1. For the Multi-label baselines, we provide also results for a multi-label model trained with a single head ("Materials" facet). For the Softmax baselines, we provide two variants: one trained with cross-entropy loss (Softmax), and another with weighted cross-entropy loss (WCE). For the hierarchical softmax models, we provide results of two variants in which losses are calculated at the node level and the tree depth level respectively. For the CLIP model, we also provide results in a zero-shot setting, without any fine-tuning. Finally, we provide results for two ensembles that combine the outputs of the multi-modal (ML), softmax (S), and CLIP models.

In Table 1 we appreciate several key insights. First, the multi-label baseline with a single head – Multi-label (M) – performs almost equally as the multi-head (Multi-label) model on the "Materials" facet. This validates our approach of extending the CNN backbone with multiple heads. Second, the Softmax baseline outperforms all other individual vision models, with

[8] https://github.com/facebookresearch/deit/blob/main/samplers.py

Table 1 Inner test set results for automatic metadata annotation on the different facets EUFCC-340K

Method	R-Precision				Acc@1				Acc@10				AvgRankPos			
	OT	M	C	S	OT	M	C	S	OT	M	C	S	OT	M	C	S
Multi-label	0.60	0.72	0.88	0.82	0.78	0.83	0.88	0.83	0.97	0.98	0.99	1.00	11.8	4.4	1.4	1.4
Multi-label (M)	–	0.72	–	–	–	0.83	–	–	–	0.97	–	–	–	5.3	–	–
Softmax	0.59	0.75	**0.92**	0.83	**0.86**	0.89	**0.93**	0.85	**0.98**	**0.99**	0.99	1.00	15.0	4.0	1.4	1.3
Softmax (WCE)	0.39	0.46	0.84	0.83	0.66	0.59	0.85	0.86	0.93	0.92	0.99	1.00	43.1	12.1	1.8	1.3
H-Softmax (levels)	0.46	0.64	0.85	0.78	0.84	0.88	0.86	0.80	**0.98**	**0.99**	0.99	1.00	81.2	13.6	2.2	1.4
H-Softmax (nodes)	0.55	0.72	0.84	0.82	0.78	**0.90**	0.86	0.84	**0.98**	**0.99**	0.99	1.00	36.4	5.5	2.1	1.3
Ensemble: ML+S	**0.66**	**0.79**	0.91	0.86	**0.86**	0.89	0.92	0.86	**0.98**	**0.99**	**1.00**	1.00	**9.7**	**3.1**	**1.3**	1.2
CLIP zero-shot	0.07	0.06	0.42	0.86	0.10	0.07	0.46	0.86	0.44	0.43	0.77	1.00	199.6	58.5	7.3	1.1
CLIP 1 tag	0.51	0.71	0.48	0.88	0.68	0.81	0.46	0.88	0.85	0.90	**1.00**	1.00	37.8	10.8	2.9	1.1
CLIP all tags	0.58	0.15	**0.92**	0.78	0.59	0.30	0.92	0.78	0.88	0.65	0.99	1.00	50.4	52.1	1.4	1.6
CLIP 1 tag prompt	0.52	0.68	0.60	**1.00**	0.69	0.76	0.58	**1.00**	0.88	0.92	0.92	1.00	41.4	10.5	3.4	**1.0**
Ens: ML+S+CLIP	**0.66**	**0.79**	0.91	0.83	0.85	0.89	0.91	0.83	0.97	**0.99**	**1.00**	1.00	13.8	3.5	**1.3**	1.3

H-Softmax stands for Hierarchical Softmax. *Facet Keys: (M) Materials, (S) Subjects, (OT) Object Type, (C) Classifications*
Bold values represent the best results, while underlined entries indicate the second-best performance

Table 2 Outer test set results for automatic metadata annotation on the different facets EUFCC-340K

Method	R-Precision				Acc@1				Acc@10				AvgRankPos			
	OT	M	C	S	OT	M	C	S	OT	M	C	S	OT	M	C	S
Multi-label	0.45	**0.85**	0.48	0.91	0.47	**0.92**	0.49	0.91	0.74	0.98	0.84	1.00	101.8	2.3	5.3	1.2
Multi-label (M)	–	0.84	–	–	–	0.91	–	–	–	0.98	–	–	–	2.5	–	–
Softmax	0.50	0.83	0.47	0.87	0.51	0.90	0.48	0.87	0.74	0.98	0.76	1.00	93.2	2.4	6.5	1.2
Softmax (WCE)	0.31	0.61	0.35	0.84	0.32	0.60	0.34	0.85	0.65	0.93	0.64	1.00	143.8	5.9	11.9	1.2
H-Softmax (levels)	0.50	0.81	0.49	0.87	0.56	0.90	0.49	0.87	0.82	**0.99**	0.72	1.00	46.9	4.1	7.3	1.3
H-Softmax (nodes)	0.42	0.84	0.48	0.82	0.44	0.90	0.47	0.83	0.76	0.97	0.72	1.00	58.5	3.1	7.4	1.3
Ensemble: ML+S	0.50	0.84	0.48	0.88	0.51	0.90	0.49	0.89	0.76	**0.99**	0.81	1.00	99.7	2.3	5.7	1.2
CLIP zero-shot	0.01	0.10	0.01	0.67	0.01	0.13	0.01	0.67	0.05	0.43	0.37	1.00	511.8	47.8	12.2	1.7
CLIP 1 tag	**0.75**	0.51	**0.87**	0.81	**0.79**	0.52	**0.88**	0.81	0.85	0.78	**0.96**	1.00	**19.1**	15.8	**2.0**	1.4
CLIP all tags	0.19	0.28	0.55	**1.00**	0.33	0.42	0.54	**1.00**	0.74	0.84	0.89	1.00	146.2	41.5	4.3	**1.0**
CLIP 1 tag prompt	0.73	0.47	0.83	0.74	0.77	0.52	0.83	0.74	**0.86**	0.83	0.95	1.00	19.6	17.1	2.2	1.4
Ens: ML+S+CLIP	0.50	0.84	0.48	0.91	0.52	0.90	0.48	0.91	0.76	**0.99**	0.93	1.00	44.9	**2.2**	4.1	1.2

H-Softmax stands for Hierarchical Softmax. *Facet Keys: (M) Materials, (S) Subjects, (OT) Object Type, (C) Classifications*
Bold values represent the best results, while underlined entries indicate the second-best performance

the Multi-label baseline performing competitively. Third, all fine-tuned multi-modal CLIP models outperform the zero-shot CLIP model. Among them, the CLIP model trained with template prompts has an overall better performance but is below the Softmax baseline in most cases. Finally, we appreciate that the ensemble methods provide an overall extra boost in all facets and metrics as expected.

In Table 2 we appreciate that the results on the Outer test set are lower in general than in the Inner set. This was expected, as the Outer test set is designed to challenge the model with data that differs from the training set (see Section 3 for details). In this setting, we appreciate that the results of CLIP models slightly outperform those of the visual models, demonstrating superior generalization. It is worth noting that in certain cases a tag label may be present in the Outer test set even though it was not included during the model's training phase, simply because it did not appear in our training dataset. In such cases, we choose to ignore this tag when calculating the evaluation metrics, as it falls outside the model's scope. The ensemble method performs poorly on the Outer test set because it averages the predictions of the multi-label and softmax methods, both of which exhibit significantly worse performance than the CLIP model in this scenario. Since averaging treats all contributing models equally, the ensemble's predictions are negatively influenced by the weaker baselines, introducing errors that the CLIP model alone would not make.

Finally, Table 3 summarizes the results of Tables 1 and 2 by showing the mean value of each metric across all facets. The key insights we extract from this summary are as follows: (1) vision-only models perform better in the Inner test set, while CLIP-based models perform better in the Outer test set; and (2) model ensembles that combine the outputs of both visual-only and multi-modal baselines provide a balanced performance across both test scenarios, at the cost of an increased computation cost.

Overall, while our models are far from achieving perfect annotation capabilities – especially on challenging edge cases– they provide significant value in assisting expert catalogers. By guiding them towards broader categories within the class hierarchy, these models enable experts to concentrate on manual annotation of the finer details, thereby alleviating the need for exhaustive scrutiny of the entire hierarchy. We further explore this idea with the design of the annotation assistant tool presented in Section 5.5.

5.4 Qualitative results

In Fig. 12, we present the qualitative outcomes for two selected images obtained using our top-performing model, which is an ensemble of the methods (ML+S+CLIP) explained earlier. Correct predictions are highlighted in green, predictions that, while not annotated in the ground truth but reasonable, are highlighted in orange, and irrelevant tags are highlighted in red. It is worth noting that the model exhibits consistency in its predictions, although ambiguities occur more often than not at the annotation level. For instance, in the right image, the prediction of the "portrait" tag for the "Subject" facet is reasonable, but not part of the ground-truth annotations of this image. This illustrates the difficulty in evaluating image tagging models for real-world applications.

5.5 An annotation assistant tool for GLAM cataloguers

To further explore the practical application of our models in real-world settings, we have developed an annotation assistant tool specifically designed for GLAM cataloguers. This

Table 3 Summary of results provided in Tables 1 and 2

Method	Inner test set				Outer test set			
	R-Prec	Acc@1	Acc@10	AvgRP	R-Prec	Acc@1	Acc@10	AvgRP
Multi-label	0.76	0.83	**0.99**	4.8	0.67	0.70	0.89	27.7
Softmax	0.77	**0.88**	**0.99**	5.4	0.67	0.69	0.87	25.8
Softmax (WCE)	0.63	0.74	0.96	14.6	0.53	0.53	0.80	40.7
H-Softmax (levels)	0.68	0.85	**0.99**	24.6	0.66	0.70	0.88	14.9
H-Softmax (nodes)	0.73	0.85	**0.99**	11.3	0.64	0.66	0.86	17.6
Ensemble: ML+S	**0.81**	**0.88**	**0.99**	**3.8**	0.67	0.70	0.89	27.2
CLIP zero-shot	0.35	0.37	0.66	66.7	0.20	0.20	0.46	143.4
CLIP 1 tag	0.64	0.71	0.94	13.2	**0.74**	**0.75**	0.90	**9.6**
CLIP all tags	0.61	0.65	0.88	26.4	0.51	0.57	0.87	48.2
CLIP 1 tag prompt	0.70	0.76	0.93	14.1	0.69	0.72	0.91	10.1
Ens: ML+S+CLIP	0.80	0.87	**0.99**	5.0	0.68	0.70	**0.92**	13.1

We show the mean value of each evaluation metric calculated across all facets. R-Prec stands for R-Precision, AvgRP for Average Rank Position
Bold values represent the best results, while underlined entries indicate the second-best performance

	
Object Types (Ground-truth): pottery **Object Types (Top-K predictions):** pottery, ...	**Object Types (Ground-truth):** coin **Object Types (Top-K predictions):** coin, ...
Materials (Ground-truth): clay, ceramic, metal, iron **Materials(Top-K predictions):** clay, ceramic, shell, combination of organic and inorganic animal material, terracotta, textile material, porcelain, rock, plant material, ...	**Materials (Ground-truth):** metal, non-ferrous metal, gold **Materials(Top-K predictions):** metal, non-ferrous metal, gold, ...
Disciplines (Ground-truth): archaeology **Disciplines (Top-K predictions):** archaeology, ...	**Disciplines (Ground-truth):** numismatics **Disciplines (Top-K predictions):** numismatics, ...
Subjects (Ground-truth): N/A **Subjects (Top-K predictions):** figure, portrait, still life, landscape, nude, marina, self-portrait	**Subjects (Ground-truth):** N/A **Subjects (Top-K predictions):** portrait, figure, landscape, still life, nude, marina, self-portrait

Fig. 12 Qualitative outcomes of the top-performing baseline on two sample images. Top-K model predictions are colored to indicate correct, incorrect, and plausible tags

tool is engineered to leverage the strengths of the ensemble model (ML+S+CLIP), providing an intuitive and interactive interface for image annotation tasks.

The core functionality of the tool is to present the hierarchical tag structure for each facet involved in the cataloguing process. Upon uploading an image, the tool dynamically updates the tree hierarchies based on the predictions from the ensemble model. Notably, it highlights the nodes within the top 10 predictions for each facet, guiding cataloguers to the most likely categories at various granularity levels. This feature is crucial, as it allows users to visually explore different branches of the class hierarchy efficiently, thereby facilitating a focused review of the relevant categories.

Figure 13 shows a screenshot of the tool in action. The image depicts how the tree hierarchies are displayed alongside the uploaded image, with the top predictions highlighted. This

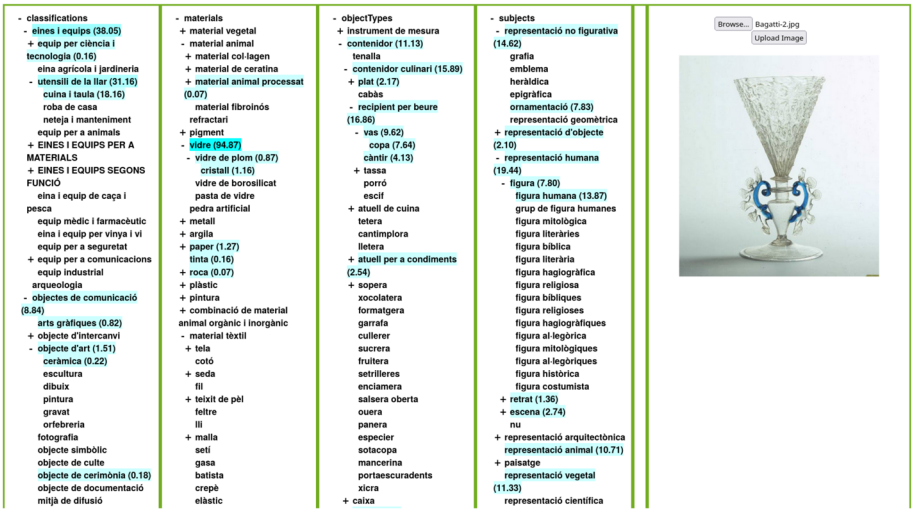


Fig. 13 Annotation Assistant Tool Interface. This screenshot displays the tool's functionality where the hierarchical class structure for each facet is shown alongside an uploaded image. The top 10 predictions are highlighted (based on their relevance score) within the hierarchies, enabling GLAM catalogers to explore various classification branches and focus on detailed annotations. Notice that branches without a highlighted relevant child can still be explored by expanding them with a mouse click on the + symbol. Note: tag labels are shown in Catalan because it is the language of the catalogers that have been testing the tool

visualization not only aids in quick navigation through the class hierarchy but also helps in understanding the model's reasoning process, making it a valuable resource for cataloguers.

The tool addresses specific pain points in the museum cataloguing process, where the sheer volume of images and the complexity of hierarchical classification can be overwhelming. By streamlining the initial stages of annotation, the tool allows cataloguers to bypass exhaustive manual review of every possible category, instead focusing on refining predictions made by the ensemble model. This approach significantly reduces the time and effort required for cataloguing, enabling museum professionals to allocate more time to contextual analysis and domain-specific expertise.

Furthermore, the tool's hierarchical visualizations provide cataloguers with a clear understanding of the relationships between predicted tags, aiding decision-making and ensuring consistency across annotations. For example, in multi-label scenarios, the tool helps to identify related categories that might otherwise be missed, such as linking an object's material and object type. These features make the tool particularly valuable for large-scale annotation projects, where efficiency is critical in order to reduce the annotation costs.

The utility of the tool has been validated by domain experts, who noted its potential to enhance productivity and support consistent metadata generation in museum workflows. By aligning advanced machine learning methods with the practical needs of GLAM professionals, this tool demonstrates the real-world applicability of our approach.

6 Conclusion

Our study into automatic metadata annotation for GLAM institutions reveals the significant potential of the EUFCC-340K dataset to enhance cataloging efficiency. By leveraging a

hierarchical annotation system, our approach addresses the long-tail distribution challenge and incomplete annotations common in cultural heritage assets. Experimental results confirm the efficacy of our baseline models, particularly when using an ensemble method. Although the performance on the Outer test set indicates challenges in generalization, our study lays a foundational framework for future research aimed at further improving annotation aid tools in the GLAM domain. In hopes of fostering further research within the community, we have made the EUFCC-340K dataset publicly available at https://github.com/cesc47/EUFCC-340K.[9] [10]

Author Contributions All authors contributed to the material preparation, study conception, and design. Data collection and analysis were performed by Francesc Net, Marc Folia, and Lluis Gomez. Francesc Net and Lluis Gomez were responsible for the baseline models' design and implementation. All authors actively participated in the analysis and discussion of the results, ensuring a comprehensive evaluation of the study findings. The first draft of the manuscript was written by Francesc Net, and all authors contributed to drafting and revising the manuscript. All authors have read and approved the final manuscript.

Funding Open Access Funding provided by Universitat Autonoma de Barcelona. This work has been supported by the ACCIO INNOTEC 2021 project Coeli-IA (ACE034/21/000084), and the CERCA Programme / Generalitat de Catalunya. Lluis Gomez is funded by the Ramon y Cajal research fellowship RYC2020-030777-I / AEI / 10.13039/501100011033.

Data Availability The dataset generated during and analyzed during the current study is available in the EUFCC-340K repository, accessible via https://github.com/cesc47/EUFCC-340K. This dataset is made publicly available to foster further research within the community. The dataset has been collected using the Europeana REST API (https://pro.europeana.eu/page/intro) and includes only records tagged with "REUSABILITY:open," ensuring the dataset comprises images suitable for open research and application (https://pro.europeana.eu/post/can-i-use-it).

Declarations

Conflicts of Interest/Competing Interests The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work was supported by the ACCIO INNOTEC 2021 project Coeli-IA (ACE034/21/000084) and the CERCA Programme/Generalitat de Catalunya. Lluis Gomez is funded by the Ramon y Cajal research fellowship RYC2020-030777-I / AEI / 10.13039/501100011033.

References

1. Van Horn G, Perona P (2017) The devil is in the tails: Fine-grained classification in the wild. arXiv:1709.01450
2. Wu M, Brandhorst H, Marinescu M-C, Lopez JM, Hlava M, Busch J (2023) Automated metadata annotation: What is and is not possible with machine learning. Data Intell 5(1):122–138

[9] https://pro.europeana.eu/page/intro

[10] https://pro.europeana.eu/post/can-i-use-it

3. Veit A, Nickel M, Belongie S, Maaten L (2018) Separating self-expression and visual content in hashtag supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5919–5927
4. Getty Research Institute (2024) Art & Architecture Thesaurus (AAT). [Online; accessed 16-Nov-2024]. https://www.getty.edu/research/tools/vocabularies/aat/
5. Torralba A, Fergus R, Freeman WT (2008) 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Trans Pattern Anal Mach Intell 30(11):1958–1970
6. Fellbaum C (1998) WordNet: An electronic lexical database. MIT Press
7. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. IEEE
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
9. Seo Y, Shin K-S (2019) Hierarchical convolutional neural networks for fashion image classification. Expert Syst Appl 116:328–339
10. Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms
11. Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Kolesnikov A et al (2020) The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. Int J Comput Vis 128(7):1956–1981
12. Van Horn G, Mac Aodha O, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S (2018) The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8769–8778
13. Van Horn G, Cole E, Beery S, Wilber K, Belongie S, Mac Aodha O (2021) Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12884–12893
14. Lutio R, Park JY, Watson KA, D'Aronco S, Wegner JD, Wieringa JJ, Tulig M, Pyle RL, Gallaher TJ, Brown G et al (2022) The herbarium 2021 half-earth challenge dataset and machine learning competition. Front Plant Sci 12:3320
15. Parmar G, Zhang R, Zhu J-Y (2022) On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11410–11420
16. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
17. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S (2022) A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11976–11986
18. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision, pp 843–852
19. Joulin A, Van Der Maaten L, Jabri A, Vasilache N (2016) Learning visual features from large weakly supervised data. In: Computer vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, pp 67–84. Springer
20. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
21. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PMLR
22. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, vol 30