Learning and Reasoning for Cultural Metadata Quality

Anna Bobasheva, Fabien Gandon, Frédéric Precioso

Learning and Reasoning for Cultural Metadata Quality

Coupling symbolic AI and Machine Learning over a semantic Web knowledge graph to support museum curators in improving the quality of cultural metadata and information retrieval

ANNA, BOBASHEVA[1]

Université Côte d'Azur, Inria, CNRS, I3S, France

FABIEN, GANDON

Université Côte d'Azur, Inria, CNRS, I3S, France

FREDERIC, PRECIOSO

Université Côte d'Azur, CNRS, Inria, I3S, France

This work combines semantic reasoning and machine learning to create tools that allow curators of the visual art collections to identify and correct the annotations of the artwork as well as to improve the relevance of the content-based search results in these collections. The research is based on the Joconde database maintained by French Ministry of Culture that contains illustrated artwork records from main French public and private museums representing archeological objects, decorative arts, fine arts, historical and scientific documents, etc. The Joconde database includes semantic metadata that describes properties of the artworks and their content. The developed methods create a data pipeline that processes metadata, trains a Convolutional Neural Network image classification model, makes prediction for the entire collection and expands the metadata to be the base for the SPARQL search queries. We developed a set of such queries to identify noise and silence in the human annotations and to search image content with results ranked according to the relevance of the objects quantified by the prediction score provided by the deep learning model. We also developed methods to discover new contextual relationships between the concepts in the metadata by analyzing the contrast between the concepts similarities in the Joconde's semantic model and other vocabularies and we tried to improve the model prediction scores based on the semantic relations. Our results show that cross-fertilization between symbolic AI and machine learning can indeed provide the tools to address the challenges of the museum curators work describing the artwork pieces and searching for the relevant images.

CSS CONCEPTS • Information systems~World Wide Web~Web data description languages~Semantic web description languages • Computing methodologies~Machine learning~Machine learning approaches • Computing methodologies~Artificial intelligence

Additional Keywords and Phrases: machine learning, image recognition, semantic Web, knowledge graph

ACM Reference Format:

Anna Bobasheva, Fabien Gandon, and Frederic Precioso. 2021. Learning and Reasoning for Cultural Metadata Quality.

1. INTRODUCTION

The value for cultural institutions lies not only in their collections but also in the knowledge extracted by curators from the artwork pieces in these collections. However, each artwork piece is peculiar by its content, by the material used, and by its style, thus the knowledge extracted for each piece has to be very precise and technical.

1556-4673/2021/12 - ArticleNumber $15.00

http://dx.doi.org/10.1145/3485844.

This increases the difficulty to define a common knowledge representation for an entire artwork collection and leads to incomplete metadata or non-uniform metadata due to the variations of annotations between several curators. In order to retrieve information in the collections, cultural institutions have put a lot of effort in the last two decades to design Content-Based Information Retrieval (CBIR) systems. However, these CBIR search engines were exploiting the latest advances of the time on visual content classification and on metadata text analysis. An example of the metadata considered at the time was the filenames or topics from the visual content. If some metadata were erroneous or missing, CBIR systems were not able to detect it as they were relying a lot on the quality of the annotation data. As far as we know, none of the systems so far were combining visual information retrieval with semantic analysis and reasoning on the metadata. The only initiative considering both views was the CLAROS project [32] but visual content and semantic web data were exploited independently by different components and not jointly to correct or refine the content. Moreover, such systems were mostly designed to provide new tools for cultural heritage content exploration by a large audience and non-experts rather than to support curators in refining and maintaining the knowledge associated with the collections.

French Ministry of Culture has a long-standing interest in the development of a reliable automated search system where artworks (e.g., paintings, sculptures) could be searched by the topics or objects represented in the image. For example, find all images with a horse in them. Currently, as illustrated on the Figure 1 the top 3 search results for the term *cheval* (horse) are the images where a horse is either on the background or partially visible and not a major subject of the artwork. Considering the number of artifacts to be annotated and the complexity of image analysis, not all of the museums have been able to describe the objects depicted with precision. Querying the images using text search on descriptors is not straightforward and can produce a number of non-relevant results or miss very relevant ones. Additionally, the existing metadata describing the art objects can be noisy or incomplete. The MonaLIA project was initiated to improve the accuracy and the details of the artwork metadata and thus support the expert work of the curators, through cross-fertilization of recent advances in fine-grained object recognition and semantic Web annotation.

Specifically, our research aims to assess whether the coupling of machine learning approaches with knowledge representation and reasoning approaches (Semantic Web and linked data) has the potential to (1) enhance metadata, (2) automate or semi-automate artwork annotation, (3) rank search results by visual relevance of a search criteria, (4) and assess the usability of an existing thesaurus for the latest AI methods.

To evaluate these questions, we experiment with the Joconde database[2][3] maintained by French Ministry of Culture that contains over 400,000 illustrated artwork records from main French public and private museums representing archeological objects, decorative arts, fine arts, historical and scientific documents, etc. The Joconde database includes semantic metadata built in the preceding project JocondeLab[4] [16, 17] developed by the Research and Innovation Institute (IRI). The iconographic description metadata was formalized in Semantic Web [14] formalism and by linking these annotations to the iconographic Garnier Thesaurus [15] and DBpedia.fr for describing the image.

The rest of the paper is organized as follows: in Section 2 we survey the related works in terms of AI approaches to improve the quality and search in cultural collections and image collections. In Section 3 we present the dataset, data processing pipeline and model selection process as well as the overall architecture

[2] http://www2.culture.gouv.fr/documentation/joconde/fr/pres.htm
[3] http://www2.culture.gouv.fr/documentation/joconde/fr/mentions_legales.htm
[4] http://jocondelab.iri-research.org/jocondelab

we designed to combine knowledge representation methods and machine learning methods in the management of a single dataset documenting a large cultural collection. Then in Section 4 we describe our approach in detection of the silence and noise and in ranking the content-based search results by visual relevance. In Section 5 we describe our approach for discovering missing semantic relations by contrasting the vocabularies and statistics of metadata. Finally, we conclude with a discussion of possible future work.

Figure 1: Search result for the images of *cheval* (horse) in the Joconde database on the Open Heritage Platform (*POP: la plateforme ouverte du patrimoine*) maintained by the French Ministry of Culture.

2. RELATED WORK

As aforementioned several search engines have been designed in the last two decades to explore cultural heritage databases, but all these solutions were either focusing on visual content analysis (just to name a few among many works, M4ART [37], Retin-3D [34], CB3DR [35]), or on semantic web data analysis (e.g. Europeana [33], MultimediaN E-Culture demonstrator [36]), and none of them were considering these two views of the content. CLAROS Project was offering to search in a database either based on knowledge representation or based on visual content but not jointly. Furthermore, most of these works have been proposed before the advent of Deep Learning (DL), which revolutionized visual content analysis.

Majority of methodological studies that use DL for image classification are conducted on standard datasets such as ImageNet[5] [1] and MS COCO[6] [2]. Real-world domain datasets popular in image classification are from biology or medicine [3, 4]. A few studies were conducted on real-world visual art images [5, 6]. Several articles have discussed the problem of art style, genre, artist and art period recognition in visual arts using CNN models [7, 8, 9] on emerging art images datasets such as the one included in WikiData [7] or the one of the Web Gallery of Art[7] (WGA) [10]. Many applications of these neural networks and transfer learning were published over the years and span over different domain-specific datasets [25, 26, 27, 28]. The advantages of using standard datasets include verified labelling [1, 2] and depiction of prominent objects captured in photographs. In the

[5] http://www.image-net.org/
[6] https://cocodataset.org/#home
[7] http://www.wikiart.org/

analysis of biomedical datasets, the objective is often to detect a specific category (class, label) with the highest accuracy [4] rather than detection of multiple often unrelated concepts. For such studies, binary and multi-class classification is common [3, 4, 5, 6, 7, 8, 9]. Art images typically have rich content with simultaneous occurrence of multiple visually relevant objects, e.g., horses, dogs, and soldiers all identifiable on a single painting.

Combining semantic reasoning and deep learning has recently become a focus of interest to try to match the implicit knowledge captured by deep networks from data with the explicit organization of known concepts in knowledge representation of these same data. This interest spreads between domain specific areas such as medicine [49, 50], and more general research [51, 52, 53].

Most of the works in this direction have only focused on very direct and basic knowledge relations such as *subClassOf* or relations between labels [39, 40, 41]. In [38], Castellano et al use convolutional neural networks to extract visual content representation of paintings. Then they build a graph between artists through a knowledge discovery process based on the visual similarity of the artworks in the feature space. This allows them to propose a new way for exploring influences between painters. However, their knowledge graph is built upon feature space similarities without any link to any knowledge base of the artworks, painters, style, etc. They infer possible painter relations from visual feature space similarities but do not reason on the knowledge they discover by relating it with confirmed knowledge from Wikipedia or other knowledge bases.

In our work, we combine two technologies: Deep Learning from images and Semantic Reasoning from structured metadata applied to visual art. We explore state-of-the-art CNN models: VGG [12] and Inception v3 [13], in a transfer learning context. Among the multiple Semantic Web development tools[8] for this project's semantic reasoning we chose CORESE[9] (Conceptual Resource Search Engine) for its ability to process RDF, RDFS, OWL SPARQL 1.1 (query and update) and for its extensions of the standards including rules and semantic distances [29]. Its standalone interface and the in-memory implementation also helped us test and prototype several queries and manipulations we will mention later.

The problem of quality of the metadata is closely related to data quality in general has been addressed before. However, some works are focused on RDF datasets and ontologies in general [20] or consistency of the SKOS-specific properties [21]. These works focus on quality problems with regards to the metadata models (RDFS, OWL, SKOS) and the linked data principles but not on domain-dependent quality checks.

Ontology-based image retrieval solutions have been proposed before [22, 23, 24]. These works have proven that with ontology-based annotations images can be found more accurately. However, the proposed approaches either employ the semantic reasoning only [22, 24] or while introducing image classifiers into the process do not utilize the Deep Learning models [23]. These contributions leverage semantics in improving the search results, for instance by adding results obtained with the transitive closure of the subtype or sub-concept hierarchy. In our project, we have a unique real-world dataset of visual art images and extensive structured but imperfect metadata that allows us to explore the combination of semantic reasoning and image recognition in order to improve both the semantic annotation and content-based image retrieval, and support curators' work.

[8] https://www.w3.org/2001/sw/wiki/Tools
[9] https://project.inria.fr/corese/

3. INTEGRATION OF SEMANTIC AND LEARNING PROCESSING OF CULTURAL LINKED DATA

We now detail the dataset, data processing pipeline, model selection process and the overall architecture we designed to combine knowledge representation and machine learning methods in the management of a single dataset documenting a large cultural collection. This is the keystone of our work as it enables us to combine and contrast results from reasoning on the symbolic metadata (RDF, SKOS) and learning on the sub-symbolic data (images) as detailed in sections 4 and 5 of this article.

3.1 The dataset

A snapshot of the Joconde database (Joconde dataset) was extracted for us in 2018. The dataset contains artwork records metadata and thesauri (Joconde KB) as well as the collection of image files. The metadata is represented in the RDF[10] format and images are JPEG files of different sizes and resolutions.

Joconde metadata is based on the ontology developed in the preceding project JocondeLab[11] and defined in a dedicated namespace[12]. The ontology defines 76 properties of the artwork. Most of the properties describe the artwork itself such as *title, author, museum*, etc., and 2 properties describe what is depicted by the artwork. The dataset contains 483,297 artwork records. 59% (285144) of them have associated images. 56% (165800) of the images have the content annotations. In terms of coverage, 37% of the properties defined in the ontology (28) are filled over 75% of the collection and 46% (35) are filled under 25% of the collection.

One particular property *sujet représenté* (represented subject) describes the content of the image and has a zero-to-many cardinality. The values of this property are based on the specific iconographic thesaurus developed by François Garnier published in 1984 [15]. The values of the *sujet représenté* (represented subject) property are especially interesting for our work because they can be considered as labels for the image classification model training. The *sujet représenté* (represented subject) concepts (REPR thesaurus) are organized in hierarchies with 12 roots. Among all of the 32274 unique REPR concepts, only 70% (22552) are associated with the artwork images, 37% (12013) are not named entities, and 2.4% (790) are associated with more than 200 images. One research question in the MonaLIA project was the quality and suitability of this old thesaurus for the latest AI methods relying on reasoning and learning. For instance, we can notice upfront that the roots of the hierarchies may not have the same abstraction level from one sub-hierarchy to another. As a concrete example, the concept *cheval* (horse) is 8 edges removed from the root concept *la nature* (nature) while the other concept *bateau* (boat) is 4 edges removed from the root concept *transport-communications*. The deepest branches are 11 edges removed from the root. The modelling and conceptualization choices and more generally the ontological and formalization commitment made in the ontology, the thesaurus and DBpedia may have a strong impact on the efficiency of AI approaches especially when then have not been designed with that purpose in mind and when they are reused from other application scenarios.

In order to scope and ground our evaluation, the French Ministry of Culture has also provided a list of 102 most searched concepts. Inspired by Large Scale Visual Recognition Challenge (ILSVRC) practices and conducting our own experiments we have concluded that ~1000 would be enough images containing a representation of a concept to train the DL algorithm. Unfortunately, and for natural reasons only 40 out of 102

[10] https://www.w3.org/RDF
[11] http://jocondelab.iri-research.org/jocondelab
[12] http://jocondelab.iri-research.org/ns/jocondelab/

concepts of interest have enough images in our dataset. In Table S5 in the Supplementary materials we provide a summary of the concepts with concepts having enough images in bold.

3.2 Data enrichment workflow with an RDF triplestore as a pivot

As mentioned above the original Joconde dataset metadata is stored in a type of database specific to RDF data called triplestore. Triplestores provide a mechanism for the storage and retrieval of RDF graphs through semantic queries and may support other types of intelligent processing including inferences and validation for instance. Our proposition is to extend this dataset with the results of image classification and of semantic reasoning by relying on the triplestore as an integration point. In order to accomplish this, we develop two-pass dataflow for training and for scoring. For the training pass, we query the triplestore using the SPARQL language [13] to create the labeled image set for training, validation, and testing. The images and labels are selected based on criteria discussed in sections 3.3 and 4.1. Then we fine-tune the CNN model described in sections 3.4 and 4.2 on the training and validation sets and test to assess the model performance on the test set.

For the scoring pass we query again but this time with different constraints to create a dataset that we run through the fine-tuned model and obtain prediction scores for every classification category for every image as described in section 4.3 and we create new triples associating the image with prediction scores, the results are represented in RDF and are stored back in the triplestore (see example in section 4.3) to be integrated and put in use with all the other metadata. As a result, we created an extended triplestore database that allows the ontology-based image search with quantified relevance of the search term. On top of this pipeline, we can then perform analytics queries leveraging all the annotations gathered and obtained and their semantics. We first designed SPARQL queries to look for the anomalies in the annotations such as noise and silence or to search for a term with better ranked results as discussed in section 4.4.

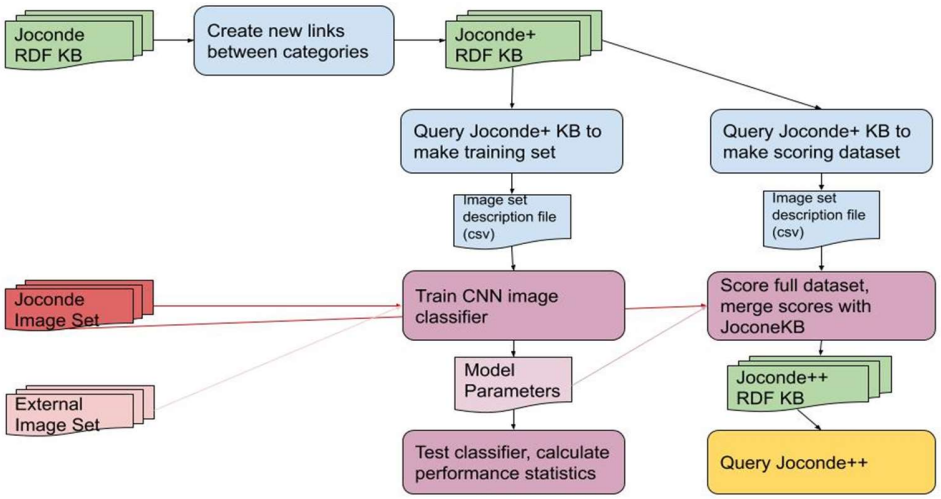


Figure 2: Data processing pipeline.

[13] https://www.w3.org/TR/sparql11-overview/

In addition, a Semantic Web triplestore provides the ability to evaluate quickly the impact of changes in the semantic models on the results of the entire pipeline.

For instance, in the early experiments, we discovered that humans on the images were not necessarily annotated by the concepts in the *être humain* (human being) hierarchy of the Garnier Thesaurus [15] and that led to a poor CNN model performance in identifying humans and animals. As shown in the examples in Table 1, on the first row's image the human is represented by concepts *soldat* (soldier) and *cavalier* (horseman) that are not sub-concepts of the *être humain* (human being) concept in the thesaurus. Also, in many instances where the *cavalier* (horseman) concept is present in the annotation the concept *cheval* (horse) is omitted like in the example on the second row.

Table 1: Examples of images and sujet représenté (represented subject) property annotation where some of the annotation concepts are missing.

Image	Property sujet représenté (represented subject)
	*figure (Révolution française de 1848, soldat, cavalier, cheval, uniforme)* (figure (French Revolution of 1848, soldier, horseman, horse, uniform))
	*scène (chasseur, cavalier, bouffon,chien)* (scene (hunter, rider, jester, dog))

To address this issue, we apply rules to create new RDF triples (Code 1) to link the concepts that are not in the same hierarchies but should be related to our tasks (search, training, etc.). To follow our example on horsemen and soldiers, we added a rule that generates RDF triple to connect the *être humain* (human being) concepts as a parent with *hiérarchie militaire* (military hierarchy) as a child which transitively connects the *soldat* (soldier), *guerrier* (warrior), *chef militaire* (military commander), etc. with *être humain* (human being).

This declarative approach and the automatic reasoning that it triggers (transitive propagation following the hierarchies) proved to be a very simple and useful technique for improving the labeling of the training and testing datasets and impact the learning performance.

CODE 1: Extract of SPARQL Query to insert new triples to the triplestore and create missing links in the thesaurus

```
insert { ?x skos:related ?y }
where {?x skos:prefLabel "cheval"@fr.     ?y skos:prefLabel "cavalier“@fr.}
```

Creating new relations between the concepts that are semantically and/or visually related but do not exist in the Garnier Thesaurus benefit the model training datasets thus creating classifiers that are more accurate.

In the example in Table 2 the second image metadata is missing the term cheval (horse) but with the link between the concepts cavalier (horseman) and cheval (horse) the image is properly labeled like the first image

3.3 Deep Learning Model

As aforementioned, Deep Learning models have revolutionized data analysis and in particular for computer vision problems such as image classification or object detection, but they require a very large training set.

Thanks to the huge number of labelled images in ImageNet dataset, Convolutional Neural Networks (CNN) are thus able to learn very efficient visual content representation outperforming all previously existing methods, even outperforming Humans, to classify images [42]. The second reason for their impact lies in their Transfer Learning capabilities. Indeed, Transfer Learning allows using a CNN that was developed and trained to solve a given problem (e.g., image classification of ImageNet dataset), to solve another but related problem (e.g., image classification of Joconde dataset) [43, 44]. The most important aspect of this capability is to be able then to benefit from a model trained on a large amount of data (usually ImageNet), then to refine or even just to use it as is for a task with few training samples [45, 46, 47]. In [48], Tan et al exploit this capability to classify images from Cultural Heritage datasets where few data are available and consequently very few training data.

Table 2: Example of the effect of the data link on the labeling.

Image	Labels & Metadata	Prediction
	*cavalier* (horseman) *cheval* (horse) *figures (cavalier, homme, cheval)* (figures (rider, man, horse)) *50440004334*	*cheval* (horse) : 0.790 cavalier (horseman) : 0.097 *voiture à attelage*(tow car) : 0.051 *chien*(dog) : 0.010 *de profil* (sideways) : 0.006 *mouton*(sheep) : 0.005 *nu*(nude) : 0.004 *draperie* (drapery) : 0.004 *épée*(sword) : 0.003 *casque* (helmet) : 0.003
	*cavalier* (horseman) *cheval* (horse) *chien* (dog) *scène (chasseur, cavalier, bouffon, chien)* (scene (hunter, rider, jester, dog)) 02110007092	cavalier (horseman) : 0.557 *cheval* (horse: 0.376 *voiture à attelage*(tow car) : 0.033 *chien*(dog): 0.012 *château* (castle) : 0.003 *drapeau* (flag) : 0.003 *maison*(house) : 0.002 *arbre* (tree) : 0.002 *casque* (helmet) : 0.002 *nuage* (cloud) : 0.002

The state-of-the-art pre-trained CNN models are typically trained as multi-class classifiers. In multi-class classification, a sample can be classified into one category among many. In an image set, one image is classified to most likely represent one of the objects from the set of multiple objects. However, the significant number of images in the Joconde dataset represent multiple objects of interest. This calls for building a multi-label classifier where a sample can be classified as a set of classes mutually non-exclusive.

At first, we have employed transfer learning from ImageNet training and then adapted to a multi-label classification by using a one-hot label encoding, a sigmoid function as output layer, and Binary Cross Entropy Loss function. To find the best performing solution for the Joconde multi-label classification we evaluated two state of the art multi class CNNs, VGG16 with batch normalization and Inception v3. Comparison details can be found in Table S1 of Supplementary materials.

The performance of these two models on the Joconde dataset is compatible while the Inception v3 is slightly faster to train and has 35 times less parameters, thus saving time and disk space. Therefore, we chose the Inception v3 model for further investigation. We have optimized other hyperparameters, such as initial learning rate, learning rate decay, optimizer algorithm and dropout rate.

Eventual training setup for a multi-label model: architecture: Inception v3, training mode: fine-tuning, dropout rate: 0.5, activation function: Sigmoid, loss function: Binary Cross Entropy, optimizer: Adam, initial learning rate: 0.001, training epochs: 20, momentum: 0.9.

3.4 Image Preprocessing for CNN consumption

The state-of-the-art CNN image classifiers take only the square images as an input. The Joconde artwork images have a wide distribution of sizes. Unfortunately, not too many square images (0.2% of all images are perfectly square, 3% are within 5% of being square). To deal with the non-square images we have tried several approaches. Filtering out "wide" and "tall" images, padding images to square, scaling and multi-cropping. Restricting the *image set* by a certain image aspect ratio may improve the training set and subsequently model performance. Empirically we found the optimal maximum for Joconde database AR = 1.2. This aspect ratio puts a serious constraint on the number of available images and the more realistic one is 1.4. Restricting the aspect ratio of the training set generally improved the performance of the classifier but, at the same time, put a constraint on the number of available images per concept. Other techniques such as padding, scaling and multi-cropping did not substantially improve model performance.

Table 3: Performance comparison for different limits on aspect ratio of the images. The experimental data: a set of images disjointly annotated with concepts *animal* and *être humain* (human being). The experimental model: Inception v3.

Aspect Ratio	Not restricted	<=2.0	<=1.4	<=1.2
F1 (macro avg)	0.77	0.84	0.85	**0.92**

In the preliminary studies the analysis of the other contributing factors on the classification results uncovered that the images of the ceramic arts are out of the distribution of the other images fooling the model and leading to a bad training. Not including this group of images into the dataset improved the performance metrics of the model up to 6%. Thus, going forward we explicitly removed the ceramic arts from the dataset, and we plan in the future to have a specific processing for this form of art. This shows another benefit of considering metadata to adapt models for very specific subsets of data.

As the results of the semantic integration of symbolic and learning processing of cultural linked data and the pre-processing performed in this first stage, we obtained a fully operational framework to start designing and evaluating methods supporting the life cycle of the data of a cultural collection and their use in search engines.

4. CHASING NOISE AND SILENCE IN METADATA AND PROVIDING RANKING TO IMPROVE SEARCH IN THE COLLECTION

The MonaLIA approach, we now propose, to improve search and metadata can be described in 4 phases: (1) we create training and test subsets images using SPARQL queries on metadata to label the images; (2) we fine-tune multi-label CNN classifier on the training set and evaluate its performance on the test set; (3) we perform prediction on the entire Joconde dataset using the trained CNN model to extend image metadata with prediction scores of classes; and (4) we perform analytical SPARQL queries on extended metadata to improve search and to chase noise and silence. As the framework of queries and model training software was developed, we gradually scaled the classifier up to 10, 20, 40 classes.

4.1 Creating training and test subsets images using SPARQL queries on metadata to label the images.

To train a CNN image classifier the labeled dataset has to be prepared. In our case, the labels correspond to the objects that the classifier is intended to identify. Therefore, in order to prepare the dataset, we need to select the images that contain the objects from the *object list* and build a subset of data containing image references and their metadata; we call it the *image set*. The selection is based on the *sujet représenté* (represented subject) property of the JocondeLab ontology (RDF properties *jcl:noticeRepr, jcl:noticeReprTerm*). These attributes contain the reference to the entities organized in a hierarchy according to the Thesaurus Iconographique published by François Garnier in 1984 [15].

For a successful model, the training image set must be balanced and large enough. From the Large-Scale Visual Recognition Challenge (ILSVRC) we know that the networks can be trained on 1000 images pre class. Also, the MonaLIA preliminary study had shown that 1000 is enough artworks images to fine-tune the pre-trained model. We also need a fraction of the images for the validation and testing.

For the training set preparation, a generic SPARQL query was developed to: (1) take an arbitrary *concept list* as an input; (2) select the artwork data records with references to the listed objects and their descendants in the hierarchy (3) take into account an optional list of exclusions of descendants that can be specified as an input in order to add flexibility in choosing the sub-concepts.

Query results are processed further to filter the records as discussed in section 3.4, make a stratified split to 3 sets for training validation and testing and save the processed result to a file to be consumed by the *data loader* for the classification step. In order to avoid copying the images to the traditional arrangement of the image datasets based on the older structure, we developed a customized *data loader* that loads the data from the Joconde image file structure provided by the Ministry of Culture.

In the multi-labeled datasets, it is very hard to balance the number of images representing each concept and each combination of the concepts increases exponentially with the increase of number of concepts ($2^C - 1$, where C is the number of concepts).

We developed a simplified method of selecting images for a training set: (a) Select all the images from the Joconde database containing the concepts as a *jcl:noticeReprTerm* property. (b) Filter out the images with more than 5 concepts from the concept list. (c) Filter out images with extremely high aspect ratio (>5). (d) Filter out the images of ceramics artworks. (e) Select 1000 images of each concept represented alone. (f) Select all the combinations of each concept with other concepts. The last step creates an unbalance that must be compensated by calculating the positive weights for each concept when calculating the loss function. Positive weights are the weights of the positive sample of the Binary Cross Entropy loss function, p_c in formula (1)[14].

$$Lc = \{l_{1,c}, \dots, l_{N,c}\}^T,\ l_{n,c} = -w_{n,c}[p_c y_{n,c} * log\ \sigma(x_{n,c}) + (1 - y_{n,c}) * log\ (1 - \sigma(x_{n,c}))] \quad (1)$$

$$p_c = \frac{count(x_{n,c}==0)}{count(x_{n,c}==1)} \quad (2)$$

Choosing the positive weights allows us to trade off recall and precision by adding weights to positive examples: pc > 1 increases the recall, pc < 1 increases the precision. We have chosen to increase the recall and calculate the pc as a ratio of negative to positive samples of the class (formula 2).

[14] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html#torch.nn.BCEWithLogitsLoss

4.2 Fine-tune multi-label CNN classifier on training set and evaluate its performance on test set.

In the preliminary study, we focused on building and assessing multi-class classifiers. For multi-class, the predicted answer is the class with the highest predicted score. For multi-label classification problems, the score threshold must be chosen to make predictions and the predicted answer is a set of classes with scores higher than the threshold(s). To make sure that the transfer learning and adaptation of the multi-class to multi-label works a set of validation runs were made on both ImageNet and Joconde images.

To evaluate model performance for multi-label classification we use *F1* score (macro average formula 3) and mean Average Precision (*mAP* formula 5). The first metric depends on the choice of the probability score threshold and the second does not.

$$F1 = \frac{1}{C}\sum_c 2 * \frac{P_c * R_c}{P_c * + R_c} \quad (3) \qquad AP = \sum_n \left(R_n - R_{\{n-1\}}\right) P_n \quad (4) \qquad mAP = \frac{1}{C}\sum_c AP_c \quad (5)$$

Where R_n and P_n are the precision and recall at the n-th threshold and C is the number of concepts.

To improve the accuracy of the classifier (*F1* metrics) the threshold selection is important. In multi-label classification especially with growing numbers of classified concepts, the *F1* metric decreases. To obtain better classification results we can vary the threshold to find an optimal recall/precision ratio.

We experimented with manual threshold validation and implemented the Proportion Cut (*PCut*) [18, 31] method that can be label-wise or global that calibrates the threshold(s) from the training data globally or per label. Label-wise *PCut* sets different thresholds for each label, which guarantees that the predicted positive rate for this label is close to the training positive rate. The more concepts we try to recognize the more beneficial it is to use calculated label-wise thresholds. The manual validation and *PCut* threshold selection shows that the higher values for the threshold (≥0.80) yield the best results.

We experimented with different number of target concepts increasing the number of classes in the outputs of the classifier. Table 4 summarizes the metrics of model performance for the different number of classes. The classes correspond to the list in the Table S1 in Supplementary materials. We have witnessed the performance degradation after the number of classes exceeds 40 because of the availability of training images. There is also an evidence that the state-of-the-art machine learning algorithm does not work as well in the domain of the diverse visual art as it does with the collections consisting of the specifically chosen photographs.

Table 4: Performance comparison for increasing number of classes. The experimental model was Inception v3.

Number of classes	Label-wise decision threshold range	F1 (macro avg)	mAP
10	0.80 - 0.90	0.72	0.80
20	0.80 - 0.95	0.61	0.65
40	0.80 - 0.95	0.52	0.54
50	0.90 - 0.95	0.46	0.47
100	0.90 - 0.95	0.26	0.28

4.3 Predict on the entire Joconde dataset with trained CNN model to extend image metadata with prediction scores of classes.

By running the model on all the images of the Joconde dataset we obtained the prediction scores for every image in the Joconde dataset and linked these scores with the artwork records by saving the scores in the same RDF format as the initial metadata using a vocabulary we designed for this purpose as shown in Code 2.

Applying the trained model to the entire Joconde database generates prediction scores for each concept for each image. We store these scores linking them with the artwork record. The association is made by creating RDF triples linking the *jcl:Notice* entity with scores. We also specify which classifier generates the scores so the result of multiple classifiers can be compared and/or used in the queries.

As a result, the RDF triplestore contains all the initial data plus all the classification results and the analysis of these results can therefore leverage semantic Web reasoning and querying capabilities in the formulation of analytic queries.

CODE 2: Extract of the prediction score represented in RDF and integrated to metadata.

```
ml:classifierRepresentedSubject a rdfs:Class ; ml:vocabID "REPR" .
ml:classifierTenClasses rdfs:label "10 classes" ;
        rdfs:comment "Classifier trained on images labeled..." ;
        rdfs:subClassOf ml:classifierRepresentedSubject .
<https://jocondelab.iri-research.org/data/notice/00000055013>
    ml:imageClassifier [ a ml:classifierTenClasses ;
    ml:detected [ a t:T523-2744 ; ml:score "0.2535"^^xsd:float ],
        [ a t:T523-1740 ; ml:score "0.2075"^^xsd:float ],
        . . .
        [ a t:T523-1051 ; ml:score "0.1096"^^xsd:float ] ] .
```

4.4 Analytic SPARQL queries on extended metadata to improve search and chase noise and silence.

To support extended metadata search on depicted concepts we developed SPARQL queries that allow the user (museum curator) to obtain the lists of images filtered by search term and its prediction scores. If the prediction score is high but the term is not mentioned in the image representation description, then it is a sign of omission (silence) in the description. If the prediction score is very low and the term is mentioned in the description, then it might be a sign that this term in the description might not be necessary (noise). The SPARQL query can also simulate a keyword search and return the results sorted by prediction scores high to low which would rank the results according to the iconographic relevance due to the fact that CNN classifiers are better in recognizing of larger objects [38, 48].

Because the multi-label classifiers recognize more than one object represented on the image, ordering and ranking the objects also allows a curator of the collection to detect silence and noise of the annotation of the image. In Tables S2-S4 in Supplementary materials we present more examples of the artwork image classification results and their impact on the quality of the information retrieval.

In Table 5 we demonstrate the top 4 search results for the concept *chien* (dog) sorted by the prediction scores in the descending order. It is clear that the deep learning model performs well when classifying the visually prominent objects and bringing the high scored images on the top of the list serves the purpose of ranking the result by relevance for a search engine or recommendation system.

To compare, on Figure 3 we present the screenshot of the results of the same search criteria returned by text-based search algorithm that is currently deployed on the portal of the French Ministry of Culture. The text-

based algorithm gives the higher rank to the results with titles that contain the search criteria, which may result in missing more relevant images if they are not properly named and annotated.

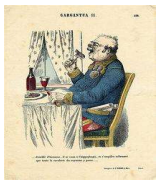
Table 5: Top 4 results of the query for concept *chien* (dog) sorted by prediction scores in descending order.

Image	Joconde Metadata	Prediction Score
	*représentation animalière (**chien**)* (animal representation (dog)) 07480003359	*chien* (dog): 1.0
	*représentation animalière (tête, **chien berger**)* (animal representation (head, shepherd dog)) 01550001084	*chien* (dog): 1.0
	*représentation animalière (tête d'animal, **chien**)* (animal representation (animal head, dog)) 50130000049	*chien* (dog): 1.0
	*représentation animalière (**chien**) ; feuillu ; fleur* (animal representation (dog); foliage; flower) 00000074825	*chien* (dog): 0.999

Figure 3: Search result for the images of *chien* (dog) in the Joconde database on the Open Heritage Platforme (POP: la plateforme ouverte du patrimoine) maintained by the French Ministry of Culture

In Table 6 we present the sample of results of querying the extended metadata to detect the noise in the existing image annotations. In these examples the query searches for the images that have the concept *cheval* (horse) or related concept *cavalier* (horseman) but have low prediction scores ($\leqslant$ 0.20). On the first row there is no visible horse on the sculpture, on the second row the horseman is barely visible on the background, and on the third row the horse is small and, although it cannot be ignored, it should have a low significance. All these examples are cases where a curator may want to revise and adjust the metadata.

Table 6: Examples of noise detection in the images that do not have a visually relevant term *cheval* (horse) with the prediction scores below 0.2.

Image	Joconde Metadata	Prediction Scores
	*cheval* (horse) *figure (saint Eloi de Noyon, évêque, en pied, bénédiction, vêtement liturgique, mitre, attribut, cheval, marteau, outil : ferronnerie)* (figure (Saint Eloi de Noyon, bishop, standing, blessing, liturgical vestment, mitre, attribute, horse, hammer, tool: ironwork)) 000SC022652	*cheval* (horse): 0.006
	*cheval* (horse) cavalier (horseman) *figures bibliques (Vierge à l'Enfant, à mi-corps, assis, Enfant Jésus : nu, livre);fond de paysage (colline, cours d'eau, barque, cavalier)* (biblical figures (Virgin and Child, half-body, seated, Child Jesus: nude, book);landscape background (hill, river, boat, horseman)) 000PE027041	*cheval* (horse): 0.009
	*cheval* (horse) *scène (satirique : Bismarck Otto von : Gargantua, repas, cheval, boisson : vin)* (scene (satire: Bismarck Otto von: Gargantua, meal, horse, drink: wine)) 5002E006121	*cheval* (horse): 0.011

In Table 7 we present the results of querying the extended metadata to specifically detect the silence in the existing image annotations. In this case the query searches for the images that have high prediction scores ($\geqslant$0.90) but no corresponding term in the metadata. In all three images the depiction of the *cheval* (horse) is prominent but the metadata is either very general (first and second rows) or has the indirect association with the horse through the concept *équestre passant* (horse riding). This method detects that the association is not captured in the thesaurus and could be an input for the linking method described in section 3.2 by providing candidate missing links for the thesaurus (e.g., horse - horse riding) and improving the results of the reasoning on the metadata.

All the examples of this section and supplemental materials showed the impact of coupling reasoning and learning on symbolic data (RDF annotations) and sub-symbolic data (images) to provide new means of improving search results and data quality, in particular noise and silence detection. To extend our ability to evaluate the quality of the metadata and their vocabulary, the next section will present a new set of metrics we proposed and experiments we performed.

Table 7: Examples of the silence detection for the concept *cheval* (horse) where the concept is visually relevant without corresponding concept in the existing annotation.

Image	Joconde Metadata	Prediction Scores
	portrait 50350012455	*cheval (horse): 0.999*
	*scène historique (guerre de siège : Lawfeld, Louis XV, Saxe maréchal de, bataille rangée)* (historical scene (siege warfare: Lawfeld, Louis XV, Saxony marshal of, row battle)) 000PE004371	*cheval (horse): 0.999*
	*figure (sainte Jeanne d'Arc, jeune fille, équestre passant, armure, casque, épée)* (figure (Saint Joan of Arc, young girl, horse riding), armor, helmet, sword)) M0301000355	*cheval (horse): 0.997*

5. DISCOVERING SEMANTIC RELATIONS CONTRASTING VOCABULARIES AND STATISTICS

Inspired by the empirical discovery of “unlinked” relations and the effect on classification performance of adding them before the labelling algorithm we studied the part of Joconde dataset metadata that pertains to the context of the image. We were looking for automated ways to discover the concepts that have a semantic relationship but are not directly linked by the predefined thesaurus. For example, concept *cavalier* (horseman) and *cheval* (horse) are semantically and visually related but are not related in the Garnier Thesaurus.

The questions we asked were: (1) Are there other pairs of concepts that can be linked to improve image context annotations? (2) Can we find them by studying pairwise context co-occurrence and/or distance between them in the knowledge graph? (3) Can we find the missing links by accessing external vocabularies including the same concepts?

In order to do this, we selected and evaluated several statistics and metrics on vocabularies both internal and external, and on the metadata of the cultural collection to identify pairs of concepts, which could be good candidates for introducing additional explicit links.

In the process we also demonstrated the interest of studying the contrast between the concepts in Garnier Thesaurus formalized in the Joconde dataset and the graph of mapped categories extracted from the DBPedia.fr showing the impact, bias and potential evolutions for the reference vocabulary selected to ground the annotations of the cultural collection.

5.1 Contrasting context similarity in image metadata with distance similarity in Garnier Thesaurus

One of the first studies we conducted was to evaluate how different were the aspects captured by the actual annotations of the images and the one suggested by the Garnier Thesaurus, in particular in terms of relations

made between the concepts. For this study, we chose the top 40 concepts that are most occurring in the image annotations among the 102 queries from the search engine provided by the French Ministry of Culture (Table S5 in Supplemental materials). We defined and formalized in SPARQL several queries to obtain the graph distance between concept pairs in the Garnier Thesaurus as well as the co-occurrences of these concepts in human annotations and in Deep Learning model predictions. We then formulated a method and devised metrics to compare the results of these different queries and identify the most related pairs of concepts accordingly and the differences we found in the proximity we got for them through each source. Intuitively, these differences capture the gap between the conceptualization as represented in the thesaurus (the theory) and the effective relations made by usage in the annotations (the practice).

The metric that we used for evaluating pairs' co-occurrence is Tversky index when α=0, β=1 in Equation 6 defines Tversky index for pairs of concepts (A, B) where A and B are the sets of annotations of artworks respectively using concepts A and B in their descriptions.

$$Tversky\ Index = \frac{|A \cap B|}{|A \cap B| + \alpha|A - B| + \beta|B - A|} = \frac{|A \cap B|}{|B|}, \quad (6)$$

We use the asymmetric Tversky Index because it is useful for determining the concept relations by giving a hint of the possible direction of the relation between the concepts, i.e., one is broader/narrower than the other.

Since the Joconde REPR thesaurus is structured as a hierarchy of concepts of the Garnier Thesaurus (narrower/broader SKOS relations), this top-down tree structure can be leveraged in metrics for evaluating path distances between pairs of concepts. To capture the level of abstractions, the edges between a concept and its broader concepts are weighted by 2^{-d} where *d* is the depth of a node from the top node. Intuitively, the lower we are in the hierarchy, the more concrete the concepts are and the less important the differences between them are. For instance, the difference between a *human* and a *building* are bigger than between a *greyhound* and a *german shepherd* which are much deeper in the hierarchy. The distance of a path between two nodes is the shortest distance across these weighted links between these two nodes through the hierarchy of concepts. This distance is formalized in Equation 8. A distance can be inverted to obtain a similarity as in Equation 9.

$$depth_w(c) = \frac{a\left(1 - r^{d(c)+1}\right)}{1 - r} - 1, \quad (7)$$

where α=1, r=0.5, and d is the number of edges from the top

$$distance_w(A, B) = depth_w(A) + depth_w(B) - 2 * depth_w\big(ancestor(A, B)\big), \quad (8)$$

where *ancestor(A,B)* is a common parent of concepts *A* and *B*.

$$similarity(A, B) = \begin{cases} \frac{2 - distance_w(A,B)}{2}, & if\ pair\ is\ connected \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

The process that we followed is: (a) Query in SPARQL the Joconde metadata to obtain co-occurrence and calculate Tversky index for each concept pair. (b) Query in SPARQL Joconde metadata and its thesaurus to obtain distances between concepts and calculate weighted distance and similarity. (c) Select the top pairs with Tversky index greater than 0.30 and sort them in descending order in a first list. (d) Rank the same pairs in a second list by the graph distance in ascending order. (e) Calculate the rank difference for each pair in the two

sorted list $rank\_difference(A,B) = rank_{similarity}(A,B) - rank_{distance}(A,B)$ (f) Finally sort the result list of the pairs by ascending order of rank difference $rank\_difference(A,B)$

As a result, at the top of the resulting list we obtain the pairs of concepts that are far away from each other or even not connected by the graph but at the same time related by context. At the bottom of the list, we have the pairs that are closely connected but do not appear together in the annotation[15].

Figure 4 represents the results comparing the ranks in the case of human annotations and predictions. Analyzing the chart, we may conclude that (1) The analysis incidentally confirms that we were correct by relating the (*cheval* (horse)*, cavalier* (horseman)) concepts during the earlier stage based on visual examination. (2) It also points out that other concepts such as (*bateau* (boat)*, mer (*sea)) and (*cheval* (horse)*, voiture à attelage* (tow car)) could be linked provided the proper predicate to capture these kinds of links.

These pairs provide candidates to improve the thesaurus and impact the labeling for image classification as well as search results. The direction of the link is such as “concept A is likely to appear with concept B”.

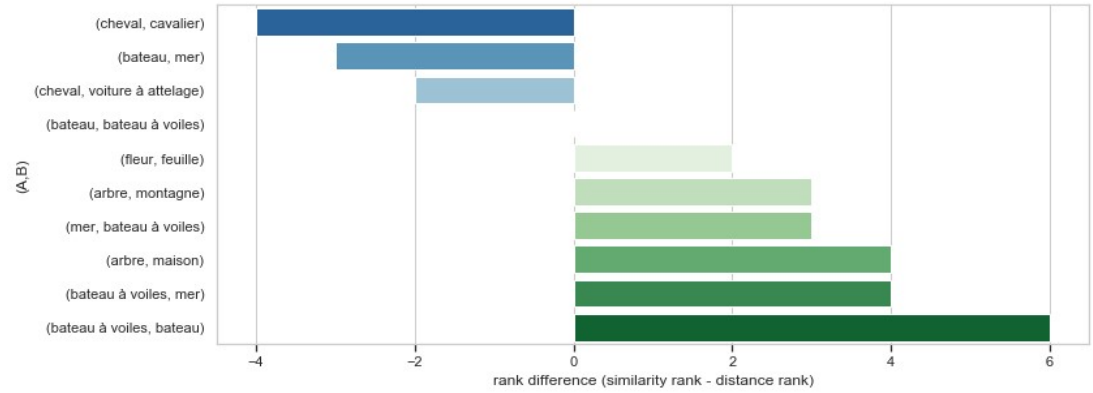


Figure 4: Low (negative and close to negative) rank difference indicates that the concept pairs that are not closely related by Garnier Thesaurus but are similar by the original annotations' context.

We applied the same process to the predictions of the deep learning models comparing the co-occurrences between the predicted concepts and distance similarity in the thesaurus and the outcome confirms the results of the comparison of the human annotation in terms of finding connections between the concepts as shown on Figure 5(a). But these results have a longer list of the pairs with a high Tversky similarity index and some of them also confirm the results discussed in the next section 5.2. For example, the visually similar “building” concepts such as *maison* (house), *château* (castle), or *tour* (tower) often classify as present at the same time on the image and it also appears that these concepts are closely related in the DBPedia.fr thesaurus as on Figure 5(b). The full figure is available in Supplementary materials (Figure S2).

[15] This may be due to the one concept being a sub-concept of another as we can see for (bateau à voile, bateau) pair

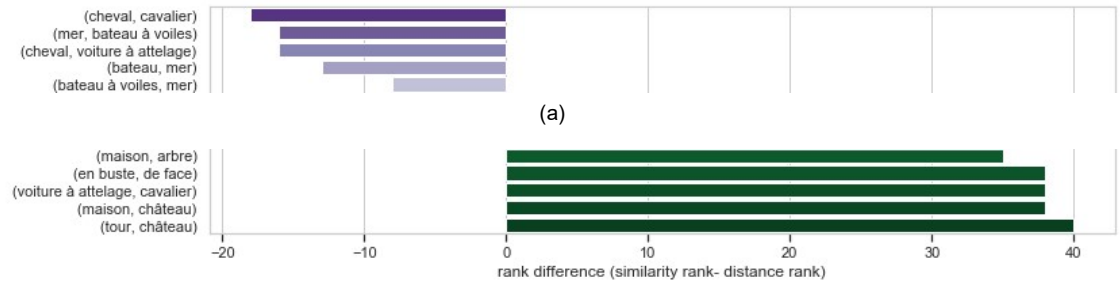


Figure 5 Low (negative and close to negative) rank difference indicates that the concepts pairs that are not closely related by Garnier Thesaurus but are similar by the predictions (score > 0.9) context.

5.2 Contrasting distance similarity in Garnier Thesaurus with graph extracted from DBPedia.fr

The previous sections and experiments all showed the impact of the structure of the thesaurus used as a reference to annotate the cultural collection on the analysis and processing we can perform on the data. We therefore decided to experiment with methods that could highlight the underlying conceptualization and formalization choices of the thesaurus itself and its bias, noises and silences. The idea was that a curator may also need to be aware of these shortcomings to improve the metadata and their use. For this study we defined and formalized a SPARQL query to obtain the graph of relations between our 40 concepts but this time in the hierarchy of the thesaurus that can be found in the linked open dataset DBPedia.fr. The hypothesis was that since the DBpedia is a much broader knowledge base it may provide some different information than a specialized cultural thesaurus.

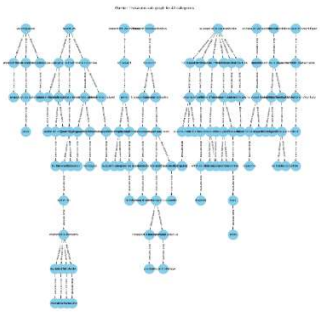
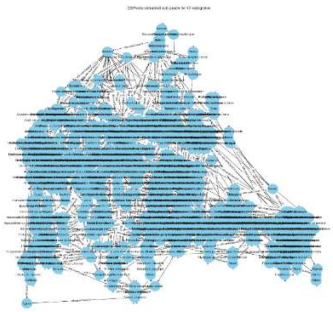
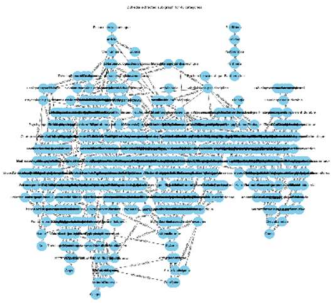
The Joconde dataset provides the mapping between the Garnier Thesaurus concepts and DBPedia.fr categories where it is possible. In the case of our 40 concepts, 3 matching categories could not be found in the DBPedia.fr (*à mi-corps* (mid-body), *de face* (front), *de profil* (sideways)) and were omitted.

DBpedia being a very dense graph, when querying the ancestors of the categories we experimented with different levels of ancestry from 3 parents to 6 parents. The graph where we query up to 6 parents of the categories becomes rooted at the top node of DBPedia.fr and therefore connects all categories in one tree-like graph. The graph with lower depth of ancestry (for example, 3) forms a forest of smaller trees and that allows for some categories not being connected as shown in Table 8.

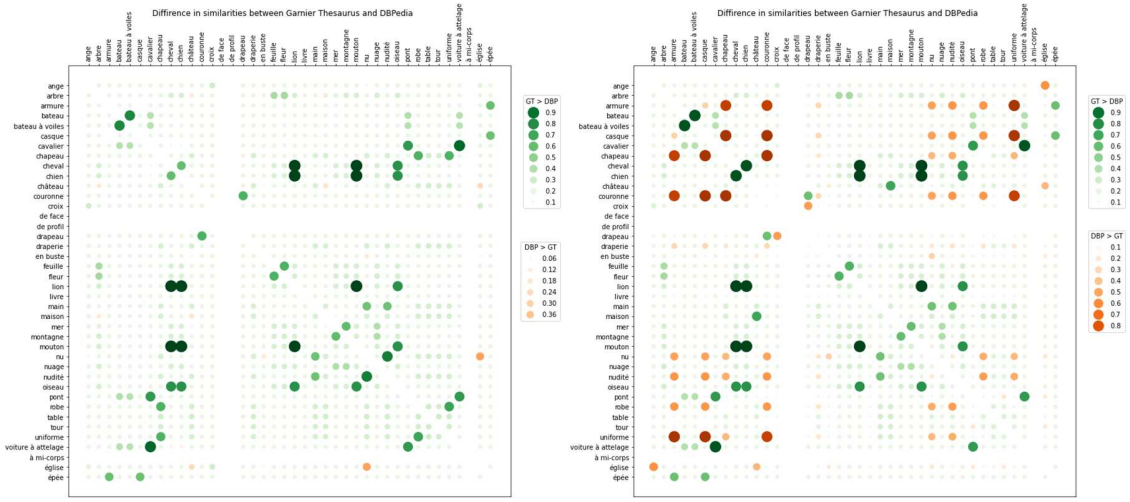
Looking at the Figure 6 (a) and considering the top-5 ranked pairs, large green circles: (1) reveals the similarity of pairs of animals in the group (*lion, mouton, cheval, chien*) that are surprisingly not closely related in the DPBedia.fr; (2) confirms the relations between concepts in pairs (*cheval*, *cavalier*) and (*cheval*, *voiture à attelage*) relations discovered by the different methods discussed in section 5.1. On the opposite side, top-5 ranked orange circles (3) reveals the pairs that have high similarity only in DBpedia and either represent the visual similarity *(château, église*) and (*château, maison),* or context similarity *(arbre, maison)* and *(en buste, nu*) and these could be suggestions of additional relations to consider in the thesaurus or metadata. We think that the artifact of the (*nu, église*) pair can be explained by the fact that often the angels and saints are portrayed

unclothed. This pair would typically be reviewed by a curator following a “human in the loop” approach and, in the end, it may not trigger an extension of the thesaurus.

Table 8: Same concepts/categories relations sub-graphs extracted from different knowledge graphs: Joconde’s REPR thesaurus based on the Garnier Thesaurus and DBPedia.fr with different ancestry depths.

Garnier Thesaurus graph of 40 leaf concepts and their ancestors	DBPedia.fr graph of 37 leaf categories and their ancestors with depth 6	DBPedia.fr graph of 37 categories and their ancestors with depth 3
		

Even more consistent results in terms of representation of visual similarity are shown by the graph extracted with less ancestors as on Figure 6 (b). It shows that the control of the level of abstraction may help in identifying relevant additions to make to the knowledge graph. Full resolution charts are included in Supplementary materials (Figures S3 & S4).



(a) (b)

Figure 6: The bubble chart of the similarity differences between the sub-graphs extracted from Garnier Thesaurus and DBPedia.fr with ancestry depth 6 (a) and 3 (b). Large circles represent large differences between the similarities.

In contrast with Granier Thesaurus DBPedias.fr sub-graph with ancestry depth 3 reveals close relations between the head covering concepts *(casque, chapeau)* (helmet, hat)*, (casque, couronne)* (helmet, crown)*,* and *(chapeau, couronne)* (hat, crown) which are visually similar. The pairs (*armure, uniforme)* (armor, uniform) *and (casque, uniforme)* (helmet, uniform) represent more the context-based similarity.

By examining the results, we can see that such comparison of the different graphs can uncover the information that can help the art collection curators improve the knowledge graph annotating its collection by adding relations, classes to the thesaurus and by considering new means to acquire them in generating the descriptions.

5.3 Using semantic relations in improving classification prediction scores

The complete semantic annotation of an artwork in the collection provides a context that can help identify suspiciously present or missing concepts. Considering the results discussed in section 5.1 we wanted to explore whether a probability of appearance of one concept of a context-similar pair can improve the probability score of the second concept in the same image. For example, if a high classification probability score for a *bateau* (boat) could influence the score of the concept *mer* (sea). To achieve this, we used a logistic regression approach. The idea is to build a pairwise regression predictor of appearance of the concept A of a pair *(A, B)* based on the presence of concept B in the Joconde dataset metadata. Both dependent variable (concept A) and predictor (concept B) are binary labels. The regression estimates the log-odds (Equation 6) of observing concept *A* when concept *B* is present compared to situations when concept *B* is not present.

$$log\left(odds(A)\right) = \beta_0 + \beta_1 * B, where\ \beta_1 = log\left(\frac{odds(A|B)}{odds(A\ no\ B}\right) \qquad (6)$$

These estimates are dependent on the dataset. If a different dataset is used it might lead to a different value of β_1. Binary indicator model compares situations when concept *B* is present or not. But because machine learning models predict concepts with continuous probability scores *S(A)* and *S(B)*, we want to reuse the regression parameters to predict an adjustment of probability score of concept A of pair *(A, B)* based on a difference of probability score of concept *B* compared to a baseline $P(B)_{base}$ that concept *B* is present.

$$log\left(odds(A)_{adj}\right) = \beta_0 + \beta_1 * (S(B) - P(B)_{base}), \qquad (7)$$

where *S(B)* is a classification prediction score of *B.* $P(B)_{base}$ can be calculated as a frequency of this concept in the dataset. Thus, we consider how much the prediction score for B is higher (or lower) than the one obtained purely by chance. For the Joconde dataset we estimated $P(B)_{base}$ by direct counting of concepts in the annotations in the training set (to be noticed, that the training set is better balanced than the entire dataset).

An alternative would be to set it up at the classification threshold determined by threshold selection algorithm during classifier training. Another, less desirable alternative, is to set it up 0.5 which indicates that the label could be equally present or not present, which is a very strong assumption for a visual art dataset.

We assume that an unadjusted estimate of S(A) corresponds to the baseline probability $P(B)_{base}$ and the adjustment could be made by using the actual prediction score S(B). The adjusted odds of A thus become:

$$odds(A)_{adj} = \frac{S(A)}{1-S(A)} * e^{\beta_1 * (S(B) - P(B)_{base})} \qquad (8)$$

In case of $S(B) = P_{base}(B)$, there will be no adjustments to the score. This approach also considers that when the score of label *B* is lower than the average, it will reduce the probability of label *A*. This approach might impact recall and precision not symmetrically which can lead to subjective decisions. For example, the presence of a *boat* often implies some body of water, however the *sea* might be present on a painting without any *boat*. We thus considered an additional threshold for the lower values of $P(B)_{base}$, below which we do not consider adjustments. This modification will adjust the prediction score of concept A when the label *B* is present and will leave *S(A)* unadjusted when $P(B)_{base}$ is low.

We developed the method of adjusting the probability scores of concept A in the context-bases pair *(A, B)* and evaluated the adjustments on the test set: (a) Estimate the odds ratio from a logistic regression on the metadata of the training set. (Equation 6) (b) Estimate the baseline probability for concept *B* on the same data. (c) Calculate the adjusted prediction score for concept *A* from the estimated odd ratio coefficient and the difference between the estimated baseline and prediction score for concept *B*. (Equation 8) (d) Evaluate the adjusted prediction using standard metrics

For step (b) we experimented with different ways of determining the baseline probability of predictor concept $P(B)_{base}$:. (1) frequency estimate based on the population of image annotations (2) uniformed estimate $P(B)_{base}$ = 0.5 (3) subjective decision estimate 1: $P(B)_{base}$ is the decision threshold for the concept B calculated during the training (4) subjective decision estimate 2: $P(B)_{base}$ is the decision threshold for the concept B calculated during the training but the adjustment to S(A) applied only if S(B) > 0.5 (5) subjective decision estimate 3: $P(B)_{base}$ is the decision threshold for the concept B calculated during the training the adjustment to S(A) applied if S(B) > 0.70 (6) subjective dichotomised estimate: $P(B)_{base}$ is the decision threshold for the concept B calculated during the training. If S(B) >$P(B)_{base}$, then (S(B) - $P(B)_{base}$) = 1.0, otherwise 0.0

The results of our method application to the example of concepts (*mer, bateau)* (sea, boat) in the population of the test set after obtaining the prediction scores with 40-class multi-label Inception v3 model and considering different types of $P(B)_{base}$ estimates are shown in Table 9.

Table 9: Evaluation of adjusted prediction of concept *mer* (sea) applying different baseline probability of predictor concept *bateau* (boat).

A = mer (sea) & B = bateau (boat)	***P(bateau)$_{base}$***	***F1***	***AP***
initial model		0.54	0.54
adjustment with frequency estimated $P(B)_{base}$	0.0778	0.35	0.48
adjustment with uniformed estimate of $P(B)_{base}$	0.5	0.42	0.48
adjustment with subjective decision estimate 1 of $P(B)_{base}$	0.85	0.51	0.48
adjustment with subjective decision estimate 2 of $P(B)_{base}$	**0.85**	**0.52**	**0.53**
adjustment with subjective decision estimate 3 of $P(B)_{base}$	0.85	0.52	0.52

Although the *F1* and *AP* metrics are lower with any of the adjustment techniques, we chose the highlighted technique as the best by looking at what changes the adjustment actually made.

It turns out that the majority of the changed outcomes were missing annotation, e.g. the original human annotations did not contain the concept *mer* (sea), while the image actually contained the depiction of a sea. Thus, model prediction could be used to detect silences with a human in the loop.

With the adjustment, model prediction moved some of the predictions above the detection threshold, as in the example in Table 10 where the original prediction score was under the decision threshold for concept *mer* (sea) 0.90 but after the adjustment the concept *mer* (sea) prediction score appeared over the threshold.

In our sample of 4463 images 109 were labelled as *mer* (sea) and 300 as *bateau* (boat), with 59 samples with both. Our procedure discovered additional 19 images containing *mer* (sea) among 300 images of *bateau* where the *mer* (sea) label was absent. Thus, we increased *mer* (sea) labels by 17% (19/109) and increased joining *mer* and *bateau* images by 32%.

This improvement, however, did not come without a price. For a small number of samples (3% of all images labeled *mer* (sea) the prediction score for concept *mer* (sea) went down below the decision threshold thus creating false negatives. This loss of information was much smaller than the discovery of new labels. In an approach where we would keep humans in the loop to validate the curation these results remain very interesting in terms of automation and scaling. We could further target to learn the optimal prediction adjustment strategy by analyzing the interactions of curators with our system.

Table 10: Example of the artwork with the adjustment of prediction score of concept mer (sea) by the prediction score of concept bateau (boat).

Image	Labels & Metadata	S(mer)	S(bateau)	$S(mer)_{adj}$
	*bateau* (boat) *paysage (Le Havre, bateau à voiles, crépuscule, soleil)* (landscape (Le Havre, sailing boat, twilight, sun))	0.8985	0.9981	0.9307

6. CONCLUSIONS

The value of cultural institutions lies not only in their collections but also in the knowledge extracted by art curators from the works of art in these collections. Quality of the services that can be offered by these collections in terms of search engines, recommendation and support depends on the quality of the catalog and its metadata. Each art object is peculiar by its content, material, and style, thus the knowledge extracted for each piece must be very precise and technical. This increases the difficulty to define a common knowledge representation and annotation method for an entire artwork collection, and leads to incomplete metadata, or non-uniform metadata due to the variations of methods and the variety of actors involved. As far as we know few of the CBIR systems are combining visual information retrieval with semantic analysis and reasoning on the metadata or at least not jointly to correct or refine the content. Moreover, such systems were mostly designed to provide new tools for cultural heritage content exploration by a large audience and non-experts rather than to support curators in refining and maintaining the knowledge associated with the collections.

In this paper we showed that the coupling of machine learning approaches with knowledge representation and reasoning approaches (Semantic Web and linked data) has the potential to (1) enhance metadata, (2) automate or semi-automate artwork annotation, (3) rank search results by visual relevance of a search criteria, (4) and assess the usability of an existing thesaurus for the latest AI methods.

In the process, we showed that state-of-the-art machine learning algorithms do not work as well in the domain of the diverse visual art as it does with the collections consisting of the specifically chosen photographs. This is also true when comparing perfect annotations of standard image benchmark datasets with imperfection of

human annotations in a specific domain. However, we found some methods that may allow the museum collection curators to improve the annotations in an automated or semi-automated way.

Our research combined the explorations in the semantic reasoning over the structured metadata to enhance the image classification and in reverse to use the results of the image classification to suggest the enhancements of the metadata.

We designed a pipeline that exploits the logic in the metadata organization to enhance the labeling mechanism for the deep learning models training, training the model, making predictions, combining the results of predictions with the initial metadata, and eventually querying the extended metadata to help with tasks that may be performed on such dataset. These tasks include cleaning and enriching the metadata and performing content-based information search with better relevance. The integration relies on a semantic Web formalization and an extension of the Joconde metadata with a vocabulary for Deep Learning model prediction scores to be used in analytical queries leveraging reasoning and querying capabilities of the RDF galaxy.

In the process we also discovered that the industry standard iconographic thesaurus may not be sufficient to describe the data for quality searches. In particular, we showed that the Garnier thesaurus was not designed for the tasks we targeted, and we therefore proposed methods to make suggestions on how to extend the metadata and the thesaurus to make it more adequate to the tasks of searching the collection and improving the catalog metadata.

For the future we consider several extensions of this work such as synthetic image generation for under-represented concepts, increasing the metadata exploited by the machine learning methods, and evaluation some of the latest approaches in segmentation for relevance scoring and further hybridizing methods between neural networks and ontology-based representations.

Combining deep learning and reasoning to improve information retrieval and predictive modelling results is a trend in AI in many areas of science such as medicine, biology, geology to name a few, where both structured and unstructured data is available. Although our work was focused on the cultural domain and a specific dataset, our proposed methodology can be applied to other datasets that include signal data (e.g. images) and structured metadata. Other domains and datasets may call for different reasoning rules and classification models, however, the general pipeline of enhancing metadata, selecting and training a classifier, and then using classification results as an extension of the metadata can easily be applied to other annotated multimedia data collections.

ACKNOWLEDGMENTS

We would like to thank the French Ministry of Culture for funding this work through the MonaLIA project and Bertrand Sajus and Laurent Manœuvre for providing the Joconde dataset and domain expert support from the cultural institution point of view.

This work is also partially supported by European Union´s Horizon 2020 research and innovation program under grant number 951911 - AI4Media

REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. International journal of computer vision115, 3 (2015), 211–252.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755.

[3] Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. Annual review of biomedical engineering19 (2017), 221–248.

[4] Yuval Mednikov, Sapir Nehemia, Bin Zheng, Oshra Benzaquen, and Dror Lederman. 2018. Transfer representation learning using Inception-V3 for the detection of masses in mammography. In2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2587–2590.

[5] Catherine Sandoval Rodriguez, Margaret Lech, and Elena Pirogova. 2018. Classification of style in fine-art paintings using transfer learning and weighted image patches. In 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, 1–7.

[6] Catherine Sandoval, Elena Pirogova, and Margaret Lech.2019. Two-stage deep learning approach to the classification of fine-art paintings. IEEE Access 7 (2019), 41770-41781.

[7] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2014. Recognizing Image Style. Proceedings of the British Machine Vision Conference 2014 (2014). https://doi.org/10.5244/c.28.122

[8] Adrian Lecoutre, Benjamin Negrevergne, and Florian Yger. 2017. Recognizing art style automatically in painting with deep learning. In Asian conference on machine learning. PMLR, 327–342.

[9] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. 2018. Fine-tuning convolutional neural networks for fine art classification. Expert Systems with Applications 114 (2018), 107-118.

[10] Benoit Seguin, Carlotta Striolo, and Frederic Kaplan. 2016. Visual link retrieval in a database of paintings. In European Conference on Computer Vision., Springer, 753-767.

[11] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. Pattern recognition 37, 9 (2004), 1757-1771.

[12] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2818-2826.

[14] Tim Berners-Lee, James Hendler and Ora Lassila. 2001. The semantic web. Scientific American 284, 5 (2001), 34-43.

[15] François Garnier. 1984. Thesaurus iconographique. Système descriptif des représentations, Paris : CNRS (1984).

[16] Laurent Manœuvre. 2015. Numérique et nouveaux enjeux pour le patrimoine. La Lettre de l'OCIM. Musées, Patrimoine et Culture scientifiques et techniques 162 (2015), 63-65.

[17] Bertrand Sajus and Marie-Véronique Leroi. 2016. Le développement du web des données culturelles. I2D Information, donnees documents 53, 2 (2016), 46-47.

[18] Jesse Read. 2010. Scalable multi-label classification. Ph.D. Dissertation. University of Waikato.

[19] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. 2005. SKOS core: simple knowledge organisation for the web. In International Conference on Dublin Core and Metadata Applications, 3-10.

[20] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. 2010. Weaving the pedantic web. In LDOW.

[21] Christian Mader, Bernhard Haslhofer, and Antoine Isaac. 2012. Finding quality issues in SKOS vocabularies. In International Conference on Theory and Practice of Digital Libraries. Springer, 222-233.

[22] Hua Chen, Antoine Trouve, Kazuaki J. Murakami, and Akira Fukuda. 2017. An intelligent annotation-based image retrieval system based on RDF descriptions. Computers & Electrical Engineering 58 (2017), 537-550.

[23] Umar Manzoor, Mohammed A. Balubaid, Bassam Zafar, Hafsa Umar, and M. Shoaib Khan. 2015. Semantic image retrieval: An ontology based approach. International Journal of Advanced Research in Artificial Intelligence 4, 4 (2015), 1-8.

[24] Yi-Hui Chen, Eric Jui-Lin Lu, and Sheng-Chia Lin. 2020. Ontology-based Dynamic Semantic Annotation for Social Image Retrieval. In 2020 21st IEEE International Conference on Mobile Data Management (MDM). IEEE, 337-341.

[25] Ruizhe Zhao, Ho-Cheung Ng, Wayne Luk, and Xinyu Niu. 2018. Towards efficient convolutional neural network for domain-specific applications on FPGA. In 2018 28th International Conference on Field Programmable Logic and Applications (FPL). IEEE, 147-1477.

[26] Taranjit Kaur and Tapan Kumar Gandhi. 2019. Automated brain image classification based on VGG-16 and transfer learning. In 2019 International Conference on Information Technology (ICIT). IEEE, 94-98.

[27] Yassir Benhammou, Siham Tabik, Boujemâa Achchab, and Francisco Herrera. 2018. A first study exploring the performance of the state-of-the art CNN model in the problem of breast cancer. In Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications. 1-6.

[28] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4109-4118.

[29] Olivier Corby, Rose Dieng-Kuntz, and Catherine Faron Zucker. 2004. Querying the semantic web with corese search engine. In European conference on artificial intelligence.

[30] Olivier Corby, Rose Dieng-Kuntz, Catherine Faron Zucker, and Fabien Gandon. 2006. Ontology-based approximate query processing for searching the semantic web with corese. Ph. D. Dissertation. INRIA.

[31] Reem Al-Otaibi, Peter Flach, and Meelis Kull. 2014. Multi-label classification: A comparative study on threshold selection methods. In

First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD 2014.

[32] Donna Kurtz, Greg Parker, David Shotton, Graham Klyne, Florian Schroff, Andrew Zisserman, and Yorick Wilks. 2009. "CLAROS-Bringing Classical Art to a Global Public. In 2009 5th IEEE International Conference on e-Science. IEEE, 20-27.

[33] Vivien Petras, Timothy Hill, Juliane Stiller, and Maria Gäde. 2017 Europeana–a Search Engine for Digitised Cultural Heritage Material. Datenbank-Spektrum 17, 1 (2017), 41-46.

[34] David Gorisse, Matthieu Cord, Michel Jordan, Sylvie Philipp-Foliguet, and Frédéric Precioso. 2007. 3d content-based retrieval in artwork databases. In 2007 3DTV Conference. IEEE, 1-4.

[35] Lirone Samoun, Thomas Fisichella, Diane Lingrand, Lucas Malleus, and Frederic Precioso. 2018. An Interactive Content-Based 3D Shape Retrieval System for on-Site Cultural Heritage Analysis. In 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 1043-1047.

[36] Guus Schreiber, Alia Amin, Lora Aroyo, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco vanOsenbruggen, Anna Tordai, et al. 2008. Semantic annotation and search of cultural-heritage collections: Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. Journal of Web Semantics 6, 4 (2008), 243–249.

[37] Egon L. van den Broek, Thijs Kok, Theo E. Schouten, and Eduard Hoenkamp. 2006. Multimedia for art retrieval (m4art). In Multimedia Content Analysis, Management, and Retrieval 2006, Vol. 6073. International Society for Optics and Photonics, 60730Z.

[38] Giovanna Castellano, Eufemia Lella, and Gennaro Vessio. 2021 Visual link retrieval and knowledge discovery in painting datasets. Multimedia Tools and Applications 80, 5 (2021), 6599–6616.

[39] Zhenzhong Kuang, Zongmin Li, Tianyi Zhao, and Jianping Fan. 2017. Deep multi-task learning for large-scale image classification. In 2017 IEEE Third International Conference on Multimedia Big Data (BigMM). IEEE, 310-317.

[40] Zhenzhong Kuang, Jun Yu, Zongmin Li, Baopeng Zhang, and Jianping Fan. 2018. Integrating multi-level deep learning and concept ontology for large-scale visual recognition. Pattern Recognition 78 (2018), 198–214.

[41] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. 2016. Learning structured inference neural networks with label relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2960-2968.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision. 1026-1034.

[43] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22, 10 (2009), 1345-1359.

[44] Amr Ahmed, Kai Yu, Wei Xu, Yihong Gong, and Eric Xing. 2008. Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In European Conference on Computer Vision. Springer, 69–82.

[45] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 806–813.

[46] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1717-1724.

[47] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks Advances in Neural Information Processing Systems 27 (2014), 3320–3328.

[48] Wei Ren Tan, Chee Seng Chan, Hernán E. Aguirre, and Kiyoshi Tanaka. 2016. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In 2016 IEEE international conference on image processing (ICIP). IEEE, 3703-3707.

[49] Ke Yan, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M. Summers. 2019. Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8523-8532.

[50] Yannick L. Kergosien, and Daniel Racoceanu. 2017. Semantic knowledge for histopathological image analysis: from ontologies to processing portals and deep learning. In 13th International Conference on Medical Information Processing and Analysis, Vol. 10572. International Society for Optics and Photonics, 105721F.

[51] Casey Breen, Latifur Khan, and Arunkumar Ponnusamy. 2002. Image classification using neural networks and ontologies. In Proceedings. 13th International Workshop on Database and Expert Systems Applications. IEEE, 98-102.

[52] Vasileios Mezaris, Ioannis Kompatsiaris, and Michael G. Strintzis. 2003 An ontology approach to object-based image retrieval. In Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), Vol. 2. IEEE, II–511.

[53] Jalila Filali, Hajer Baazaoui Zghal, and Jean Martinet. 2020. Ontology-based image classification and annotation. International Journal of Pattern Recognition and Artificial Intelligence 34, 11 (2020), 2040002.