Check for updates

REVIEW

Automated metadata annotation: What is and is not possible with machine learning

Mingfang Wu[1], Hans Brandhorst[2], Maria-Cristina Marinescu[3], Joaquim More Lopez[3], Margorie Hlava[4] & Joseph Busch[5†]

[1]Australian Research Data Commons, Australian Research Data Commons, Melbourne, Australia, Australia

[2]Iconclass, Voorscoten, The Netherlands

[3]Barcelona Supercomputing Center, Barcelona, Spain

[4]Access Innovations, Albuquerque, New Mexico, USA

[5]Taxonomy Strategies, Washington, DC, USA

Keywords: Metadata annotation; Metadata, Machine learning; Culture heritage; Research data

Citation: Wu, M.F., Brandhorst, H., Marinescu, M.-C. et al.: Automated metadata annotation: What is and is not possible with machine learning. Data Intelligence 5 (2023). doi: 10.1162/dint_a_00162
Received: March 15, 2022; Revised: April 18, 2022; Accepted: June 7, 2022

Corrected Proof

ABSTRACT

Automated metadata annotation is only as good as training dataset, or rules that are available for the domain. It's important to learn what type of data content a pre-trained machine learning algorithm has been trained on to understand its limitations and potential biases. Consider what type of content is readily available to train an algorithm—what's popular and what's available. However, scholarly and historical content is often not available in consumable, homogenized, and interoperable formats at the large volume that is required for machine learning. There are exceptions such as science and medicine, where large, well documented collections are available. This paper presents the current state of automated metadata annotation in cultural heritage and research data, discusses challenges identified from use cases, and proposes solutions.

1. INTRODUCTION

Artificial Intelligence (AI) is often defined as "The simulation of human intelligence process by machines, especially computer systems" [1, 2]. This means it can be many things: statistical, rules engines, or other forms of intelligence and it encompasses several other domains including natural language processing,

† Corresponding author: Joseph Busch (Email: jbusch@taxonomystrategies.com; OICID: 0000-0003-4775-8225).

© 2022 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

computer vision, speech recognition, object recognition, and cognitive linguistics to name a few. It also intersects with cross language retrieval and automatic translations.

Machine Learning (ML), a subset of AI, trains mathematical models by learning from provided data and making predictions. The learning can be

- supervised by learning from labeled data and predicting labels for unlabeled data,
- unsupervised by discovering structure or groups in data without predefined labels, or
- semi-supervised by taking advantage of both supervised and unsupervised and learning from a small number of labeled data.

We have seen successful applications of ML, for example, image recognition for identifying objects from digital images or videos, speech recognition for translating spoken words into the text, personalized recommendation systems, and classification systems in email filtering, identification of fake news from social media, and many more. The ability of commercial companies to collect large volumes of data, labeled and unlabeled, through user interactions, is a key factor in the success of these applications.

In this paper, we examine the ML applications supporting curation and discovery of digital objects from cultural, archival, and research data catalogs. These digital objects could be images, videos, digitized historical handwritten records, scholarly papers, news articles, dataset files, etc. What these catalogs have in common is that metadata needs to be generated for describing digital objects in curation, especially descriptive metadata that provides information about the intellectual contents of a digital object [3]. Descriptive metadata includes title, description, subject, author etc.

Traditionally, curators (or data owners) tag metadata to an object. This is a labor-intensive process, and in today's digital world, it is almost impossible for curators to tag metadata to each digital object due to the massive volume of digital objects generated and historical cultural artifacts being digitized each day. There have been investigations of exploiting ML and AI methods in generating descriptive metadata [4]. For example, Maringanti et al. [5] applied a deep neural network ML model to generate captions and subject labels for historical images in a digital library. Suominen et al. [6] have trained an ensemble ML model, namely Annif, to automatically annotate subject labels to a collection of digitized paper archives from the National Library of Finland. Here labeling an object with a subject label is also a classification task in that all subject labels represent categories to classify an object.

There have also been investigations of ML based computer vision algorithms for recognizing objects in artwork [7, 8]. Convolutional Neural Network (CNN), a popular computer vision algorithm, has been applied in identifying and classifying art works. For example, Milani and Fraternali [9] developed a CNN classifier for the task of iconography class recognition. This work also re-used common knowledge extracted from a pre-trained model on ImageNet [10] to address the problem of the limited amount of labeled data in the domain. Cetinic and Lipic et al. [11] went beyond object identification and classification. They applied and trained CNN models to predicate subjective aspects of human perceptions: aesthetic evaluation, evoked sentiment, and memorability.

Corrected Proof

Although the above investigations have shown promising results, they also identified barriers for models to match to or substitute curators or human experts, largely due to: 1) A lack of commonly available high-quality annotated datasets from the studied areas; and 2) Annotating metadata is a cognitive intensive activity; it requires curators have both domain knowledge and general knowledge of the cultural and social context to situate an object within a proper context.

In this paper, we will present three use cases that discuss the application of ML for three different tasks. The first use case presents a task of not only identifying objects in an iconography, but also going further to infer what an object symbolizes; the second use case is to generate a caption for a digitized historical image; and the third use case is to annotate subject labels from the short description of a dataset. Through the use cases, we will then discuss what is and isn't possible with ML, and identify challenges for future improvement.

2. USE CASES

2.1 Use Case 1: Bridging the Gap: AI and Iconography

Perhaps the most interesting challenge for Artificial Intelligence and Machine Learning applications in the field of iconography and the subject analysis of images is to adjust the answers Information Technology developers are formulating to the questions researchers in the Humanities actually have, thus bridging the gap between technical solutions and real-life problems.

The complexities of those questions are widely divergent. Sometimes we simply want to know the source of an iconographical detail or of a composition. Sometimes we are interested in how the concept of the king as a vigilant ruler of his people could be visualized.

Even if pattern recognition algorithms are able to extract pictures of a crane, a dog, and a lion (Figure 1) from a large corpus of early modern prints, they will probably not retrieve them as hits for a query for symbols of "*good government*." It is unlikely that this type of high-level interpretation will be available anytime soon.

Figure 1. Examples of iconography.

Corrected Proof

It seems a good idea, therefore, to test the state of affairs in the field of AI and ML against a more concrete type of research question, and feed pattern recognition algorithms some fairly straightforward patterns which they can then attempt to detect in a large corpus of images; for example, the patterns that are all linked to the human body, but, more importantly, are all derived from actual research projects. As shown in Figure 2, the first pattern is that of a man or woman pinching his or her nose, usually as a response to the perception of a disagreeable smell[①]. The second pattern is that of the posture of crossed legs in portraiture or portrait-like representations of men and women[②]—possibly as an expression of self-confidence or authority. The third pattern is that of the extended hand, often as a gesture of speech or command[③].

Figure 2. Examples of postures in art work.

Since all three patterns correspond to ready-made concept labels in the Iconclass system, they are—in principle—connected to rich, multilingual collections of resources that are represented by language models as "*bags of words*." The Iconclass concepts these positions of the human body can be tagged with, are as follows:

- 31B6222 · holding one's nose closed
- 31A2627(+53) · standing, leaning, sitting or lying with legs crossed (+ sitting)
- 31A2512(+9321) · arm stretched forward (+ speaking)

They are concrete and quite distinct, so detecting them in a large and varied corpus of images should present an interesting challenge. What should make it all the more rewarding is that there are actual humanities researchers with an immediate interest in the result.

① Linked to the Odeuropa project of, a.o. Lizzie Marx and Inger Leemans. See: https://odeuropa.eu/

② Linked to the research of postures and gestures in 17th century Netherlandish art, by Aagje Lybeer of Ghent University, supervised by Koenraad Jonckheere, Elsje van Kessel and Karin de Wild

③ A personal research interest of the present author.

Correction Proof

To investigate what automated—or more accurately "*computer-assisted*"—tagging of images is capable of at present, an important additional question is how large the dataset should be with which we train the algorithms. How many examples of the pattern we are researching do we need to collect and how should we prepare the images? More to the point, how do we isolate the relevant details when they cover only specific areas of the bitmaps?

Just how useful would it be if an "*uncomplicated*" gesture like that of a raised hand could be identified by a pattern matching algorithm? Over the years the staff of the *Index of Medieval Art* at Princeton[4] have manually tagged so many images with the term "hand raised" that the query algorithm does not produce all the hits when you search with these keywords. The result is limited to 10,000 hits.

A pattern matching algorithm has not been able to assume the role of the expert staff of the Index, but it is fair to assume that they would not mind if some pattern recognition software would alert them to candidate images. It might seriously speed up the cataloging process.

2.2 Use Case 2: Caption and Labels for Paintings

Metadata about cultural heritage (CH) artifacts is useful for tasks such as cataloging, search, or question answering. But it is also useful for online accessibility to art works by populations such as visually impaired users (via rich alt text). Despite this, the metadata for CH remains sparse and generally focuses on context, author, and style. There is an implicit assumption that one sees the artwork, and thus there is no need to describe its content. Visual content is nevertheless fundamental for the tasks we mention above, and it is a prerequisite to try to infer higher level knowledge such as symbols and meaning.

Trained models exist for image classification, object detection, and caption generation, and they perform very well when tested on photographs of every-day life and things. This is not the case when tested over art works for multiple reasons:

- Anachronisms: many objects present (and sometimes overrepresented) in photographs are anachronistic when considering images from past centuries (e.g., cell phones, cars, TVs, microwaves, headphones, teddy bears, etc.); objects that look like them will be identified as such, resulting in clergy holding tablet computers or cell phones, or knights riding motorcycles.
- Some objects have changed shape over time: e.g., printing devices, plows, phones, etc.
- Some actions are not depicted in modern photographs: subjects such as decapitations, killings, torture, rapes, and so on, are depicted in art but not so frequently in published photographs. A model trained over photographs will not be able to recognize these actions when presented with a painting.
- Symbols and imaginary beings: While these concepts could be rendered with modern technology, the existing models are trained over extensive datasets of pictures of everyday life, where there are no devils, saints, halos, angels, flying dragons, unicorns, etc.

[4] https://theindex.princeton.edu/

Corrected Proof

- Existing models only recognize generic classes, e.g., person, flower, bird, etc. To be able to transmit the intended (e.g., iconographic) meaning, they need to be more specific. In Figure 3 below, the image on the right would not be very instructive if described as a person on a horse.
- Variation in representation styles raises problems even for simple objects such as a book.

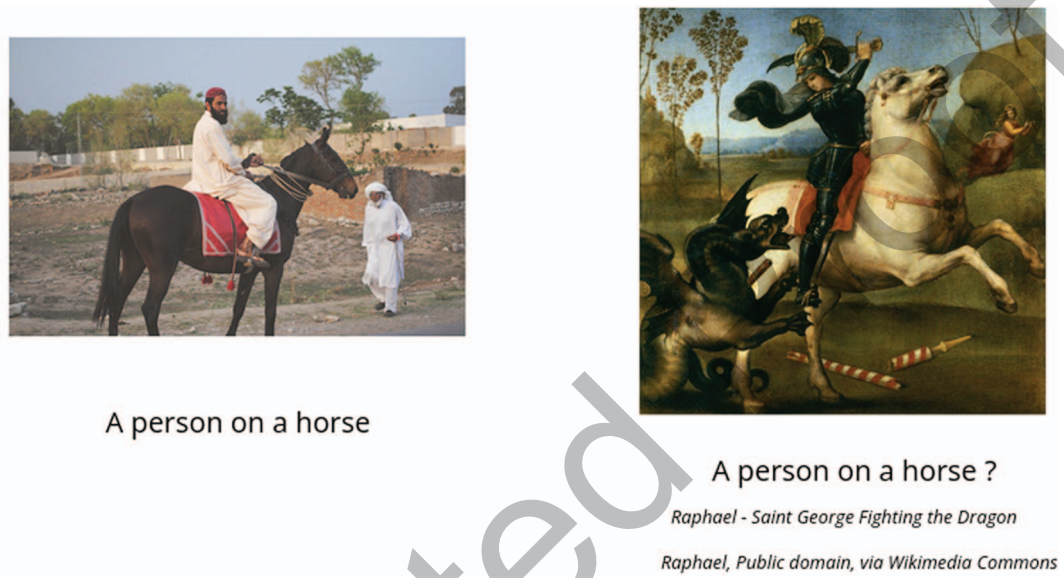


Figure 3. Two artworks showing the same relation between the main objects; while the first one is a reasonable caption, the second one is too generic and has no cultural insight.

To overcome these challenges, one needs to train models that are specific to CH over datasets containing artworks. Two problems arise: first, there are not a lot of artworks by ML/DL[5] standards; we have already discussed the second problem, which is that the existing works are usually not content-labeled. The best option, as far as we know and have experimented with, is to use transfer learning from the pre-trained photograph models and manage to label as many artworks as possible to train accurate object detectors and caption generators.

Let us look at these two tasks separately. To train an object detector, one needs a training dataset including tight, class-labeled bounding boxes for all the objects depicted in the artwork. There is no way we know of around manual annotation of images, at least until a decent model can be trained, which may then be used in a semi-supervised manner for additional images. In the next paragraphs we explain our experiments to gather a significant dataset annotated for object detection. To train a caption generator, one needs a training set of aligned image/description pairs. At the end of this section, we will describe how we perform this task in a project under course[6].

[5] DL-Deep learning, a machine learning method based on artificial neural networks with representation learning.
[6] https://saintgeorgeonabike.eu/

Corrected Proof

2.2.1 Object Detection

We train a deep learning model that predicts object bounding boxes and labels based on a dataset of manually annotated artwork images. The dataset contains 16,000 images, which we split into training and testing sets. The labels predicted by the model are predefined as part of a list of 69 classes. Some examples can be seen in Figure 4.

Figure 4. Examples of object identification via labeled bounding boxes.

As we already advanced, broad classes are not very useful when describing cultural heritage objects, especially when sets of co-occurring objects have special significance or define more complex conceptual scenes or symbols. Given that a large number of specialized classes would make it more difficult to train a good model, we approach this challenge by refining simpler object labels using a language model (for English) to generate new labels.

The bounding box labels are put in a textual context where the more precise label is predicted by the fill-in-the-blanks task performed by a model trained for mask language modeling. The textual context expresses the following: an entity with *label1* in relation to an entity with *label2* is <THE MORE PRECISE LABEL>. For example, *a person (label1)* on *a horse (label2)* is *a rider (THE MORE PRECISE LABEL)*. The language model predicts *rider* as one of the more precise labels in this context. When the bounding box labeled with the more precise label is related to other bounding boxes, this label can be further specialized.

Corrected Proof

For example, if a bounding box labeled *sword* is in a hold relationship with the bounding box previously labeled as *rider*, then the latter can be further specialized as *knight*.

The reader will have noticed that we mentioned relationships between objects. These are needed to increase the precision of predictions; it's not the same "a man on a horse" as "a man feeding a horse." One way to generate (a subset of) these relationships is heuristically, based on bounding box analysis or pose estimation.

The object detection task for CH has several limitations, of which the most important are minority classes and overlapping objects. Given the limited number of artworks, some of these limitations may not be fully overcome. This means that the role of the "man in the loop" is very important, even more so when some artworks may depict unusual scenes making them hard to predict.

2.2.2 Visual Descriptions

Another useful computer vision task for CH is the automatic generation of visual descriptions of artworks. We distinguish here two possibilities: the generation of descriptions in natural language or the generation of triples representing the objects and their relationships; together these triples form a knowledge graph.

The task of generating natural language descriptions is similar in spirit to the object detection task in that it is performed based on deep learning from an (extensive) existing dataset that contains aligned image/description pairs. The crux of the problem is that such aligned datasets do not exist and the question is how to best obtain them. The traditional approach in such cases is to get at least five manual descriptions for each image, typically via a crowdsourcing campaign. One of the fundamental issues here is the quality of the descriptions, and how to evaluate this automatically, if possible. We are experimenting with an alternative approach based on training a sentence classifier. The intention is that this classifier will be able to select, from complex and mixed content/context/style descriptions, those phrases that are relevant for the visual content on an image; simply put, it will be able to discriminate visually descriptive phrases from those that refer to other information. For example, in the description of Ribera's painting *Jacob's Dream* in the Prado Museum collection[7] (Figure 5), only three sentences refer to the content depicted in the painting. The classifier is able to detect them. Figure 6 shows the prediction of one sentence as descriptive (DESC), that is, a sentence that describes the content of the painting, and one non-descriptive sentence (NODESC).

The sentences classified as DESC will be paired with the image of the painting in the aligned corpus used for training a natural language descriptor.

Instead of generating natural language descriptions, one may want to generate a knowledge graph from the semantic labels of objects and the relationships between them (as suggested in the object detection task). Very simple sentences may be generated from the triples, but they will certainly not have any of the

[7] https://www.museodelprado.es/en/the-collection/art-work/jacobs-dream/c84dbc72-af49-4a7c-88df-a66046bc88cd

Corrected Proof

This painting tells of Jacob the Patriarch´s mysterious dream, as told in Genesis. He appears asleep, lying on his left shoulder with a tree behind him. On the other side is the ladder of light, by which the angels ascend and descend. This subject demonstrates Ribera´s skill at constructing metaphoric discourse. He uses the image of a shepherd resting in the countryside to describe one of the best-known Bible stories. The foreground view of the solidly constructed figure and the scene´s realistic features bring realism to the miraculous dream described in a ray of light under a blue and gray sky. Here, Ribera offers one of the numerous proofs of his delicate sense of color and his exquisite compositional capacity, masterfully placing the foreground volumes in diagonal contraposition. This is probably one of the paintings mentioned in 1658 in the inventory of don Jerónimo de la Torre, and it remained in his family until 1718. It reappeared in 1746 among Isabel de Farnesio´s paintings, attributed to Murillo.

Figure 5. Description of Ribera's painting *Jacob's dream* in the Prado Museum collection.

```
example1 = "The man is sleeping, leaning on his left arm."
example2 = "This is one of the most well known biblical episodes."
p1 = rp.predict(example1)
print(example1, p1)
p2 = rp.predict(example2)
print(example2, p2)

The man is sleeping, leaning on his left arm. DESC
This is one of the most well known biblical episodes. NODESC
```

Figure 6. Example of automatic DESC and NODESC classification of sentences in the description of the artwork on the left (right, example 1: DESC, example 2: NODESC).

sophistication nor feel of natural language statements. This doesn't imply that they are less useful, especially for further inference, search, or question answering.

In summary, we see the limitations in applying ML in this use case, including: 1) Metadata, especially visual content, for CH is sparse, 2) The visual content in existing metadata for artworks is mixed with other information and must be automatically detected, 3) General ML needs to be fine-tuned for the studied domain where art works are anachronistic, and 4) The task involves higher level knowledge that may not be embedded in objects and so we have to learn that from external resources.

Corrected Proof

2.3 Use Case 3: Applying ML Methods in Annotating Subject Metadata to Dataset Records

This case study focuses on research data, which are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship [12]. Digital representations of research data take many forms, e.g., image, video, spreadsheets, plain text files, or databases. The curation, discovery and sharing of research data is mainly through existing metadata, especially if this is published following FAIR data principles [13]. Metadata provides information about data and plays an important role in data cataloging services to organize, archive, identify and discover datasets, as well as facilitate interoperability between data catalogs.

In this case study, we report on an experiment that applied machine learning approaches to automatically annotate metadata records from a data catalog with subject metadata [14]. Subject metadata is a topical property of a dataset, and is a powerful way of organizing knowledge and linking resources for interoperability and discovery. Subject metadata is included in almost all metadata schemas, which usually specify the subject metadata in terms that are part of a well-developed classification code or controlled vocabulary.

In this experiment, we took a collection of metadata records from an Australian national research data catalog—Research Data Australia (RDA)[8]. Metadata record providers are urged to annotate the subject metadata with category labels from the Australian and New Zealand Research Classification—Fields of Research (ANZSRC-FoR)[9], which is used widely in the Australia and New Zealand research community for reporting research outputs, among many other uses. In its 2008 version, the ANZSRC-FoR classification vocabulary has 1417 labels that are arranged in three hierarchical levels containing 22, 157, and 1238 labels as we go from the first to the third layer. This experiment used only the top level 22 labels or categories.

We collected 142,792 data records from RDA. We found that only about 55% of them (78,535) have at least one ANZSRC-FoR label, and about 93% of these records concentrated in four subject areas. In order to avoid bias toward big categories in the process of developing training models, given that the bias may negatively affect the classification accuracy of the trained models for the rare categories, we downsized big categories but still maintained the relative size between categories, for example, for the category "04" and "21" that each has over 10,000 records, we randomly chose 600 records from each category. The column "all data" in Table 1 shows the number of records per category, and the column "down size" shows the actual number of records from each category that are used in the experiment. With this dataset, we extracted words from a record's title and description, and represented each word based on their numeric value from tf*idf after applying a stopword list and the Lemmatizer stem method. We then trained four supervised machine learning classifiers: multinomial logistic regression (MLR), multinomial naive bayes (MNB), K Nearest Neighbors (KNN), and Support Vector Machine (SVM), to annotate test records with the 22 labels in the top layer. Each classifier used three quarters of the records from each category to train the models, and the remaining one quarter was used for testing. The processes of data clearing, model training and testing, and evaluation are described in [14].

[8] Research Data Australia: https://researchdata.edu.au/

[9] Australian Research Council: Classification Codes - FoR, RFCD, SEO and ANZSIC Codes (2008): https://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/4AE1B46AE2048A28CA25741800044242?opendocument

Corrected Proof

The result in Table 1 shows that the four trained classifiers have very close performance, with MLR performing slightly better than the other three. By looking at the classifiers' performance per category, we can see a big variation, with six categories achieving accuracy closer to 1, seven are under 0.5, and the rest are in between. The poorly performing categories are those with fewer records annotated even though we had downsized the big categories. Further investigation also indicated that those categories with high accuracy are represented by more distinct features than those that aren't. For example, the top five weighted features representing the category "Earth Science" (with the code 04) are *earth*, *airborne*, *geophysical*, *mount,* and *igsn*, while the category "Commerce, Management, Tourism and Services" (with the code 15) are *study, financial, survey, university,* and *dataset;* where the term *university* and *dataset* are general to many records from other categories.

Table 1. Performance of the four classification models per annotated category label.

2 digits code	MLR	SVM	KNN	MNB	down size	all data
01	0.29	0.00	0.41	0.33	*111	111
02	0.97	1.00	1.00	0.92	300	3537
03	0.73	0.61	0.60	0.59	499	499
04	0.96	0.98	0.92	0.90	600	10147
05	0.61	0.63	0.68	0.49	400	5417
06	1.00	1.00	0.64	0.96	600	24520
07	0.63	0.52	0.77	0.42	200	1032
08	0.45	0.22	0.53	0.26	*386	386
09	1.00	1.00	0.94	1.00	200	2031
10	0.29	0.00	0.20	0.00	*128	128
11	0.68	0.69	0.63	0.64	400	1409
12	0.61	0.95	0.67	0.66	*174	174
13	0.58	0.91	0.69	0.67	*148	148
14	0.41	0.00	0.58	0.57	*122	122
15	0.21	0.00	0.18	0.00	*76	76
16	0.56	0.50	0.55	0.54	300	723
17	0.40	0.00	0.32	0.67	*112	112
18	1.00	1.00	0.99	0.98	400	849
19	0.82	0.69	0.76	0.54	*343	343
20	0.89	0.85	0.26	0.81	300	553
21	0.97	0.96	0.99	0.88	600	32592
22	0.34	0.00	0.65	0.44	*79	79
micro ave	0.70	0.67	0.66	0.66	4799	84988
macro ave	0.65	0.57	0.63	0.60		
weighted ave	0.76	0.71	0.70	0.68		

A few questions arise from this use case. First, how can we effectively apply the trained models to the records in RDA, since the best performing model achieves accuracy above 0.95 only for about one third of the categories? Because the categories with low accuracy are largely due to the lack of a good number of training data points for underrepresented categories, one solution could be to apply the best model to

Corrected Proof

the best performing categories as an initial step, and re-train models when these underrepresented categories acquire more records. This would involve setting up a workflow that periodically trains and updates models as more records are ingested into RDA in the future. Another solution is to seek resources outside of the catalog for pre-trained language models. Some new technologies (e.g., deep learning and transfer learning) may provide a way, for example, we may fine-tune contextualized word embeddings from pre-trained Bidirectional Encoded Representations from Transformers (BERT) to the specific annotation task as described above. Either way, these good quality resources for training are created or at least verified by humans, which indicates that humans are still playing an important role in training ML models.

The second question is: When we operationalize the best performing model and make the recommendations visible to end users, how do we treat the category labels that were suggested by the ML model: should they be used and displayed in the same way as those labels assigned by humans? This is actually the question about how to interpret and evaluate machine generated category labels: when an ML model suggests a label, this is based on probability and represents the likelihood that a record is about the labeled topic based on evidence from the training dataset. It is important to convey this likelihood (uncertainty) information to a human user, so the user understands and can still trust and continue to use catalog functions, for example facet filters and facet search, that are based on subject labels. A catalog can go further to collect feedback from users, for example, is a label right or wrongly assigned to a record, or ask users to provide a label; this feedback may lead to improved ML intelligence and classification accuracy.

Last but not least, a question arises about the degree that the results can be generalized: Can other data catalogs apply the best performing models we (or others) have trained? It is well documented in the literature that no classifier exists that works equally well for different text classification tasks; even for the same task, the performance of an ML classification model largely depends on the nuances of training data such as class distribution, features, and data volume [15]. This indicates that each data catalog needs to train their own models. However, some resources can be shared among data catalogs, for instance word associations and training datasets with the same or similar category labels, as well as software source codes and tools like Annif [6] that are used for training. The more resource sharing occurs between the different data communities, the better (annotation) results each catalog can achieve.

3. DISCUSSION

As discussed in the Use Cases, accurately detecting themes in cultural materials and datasets is a challenge. But it's important to point out that complete and consistent tagging is as much a challenge for people as it is for algorithms. It's a problem when describing images, but it's also a problem with any metaphor-laden content, such as poetry, which is imagistic but textual. The third use case considers datasets which pose a somewhat different categorization challenge, but is no less a contextual challenge than when working with cultural heritage materials. The problem with datasets is that the text is sparse, thus not amenable to natural language processing, and any visualization usually presents a very different problem than with cultural images.

Corrected Proof

How can the categorization problem be simplified so that we are not trying to solve the general case? Methods for description and categorization have been devised for many domains that can sometimes make categorization and meaning detection easier, and can sometimes be a framework for guided if/then logic. As described in Use Case 1, ICONCLASS has been used to categorize a lot of art works. An iconographic ontology such as the ICONCLASS schema has been widely applied to image collections for inferring meaning and storylines of complex images. These well categorized examples could be used as training datasets for automated methods to guide detection of metaphoric and symbolic imagery. Also, the ICONCLASS scheme, which is an enumerated hierarchical classification, could be used to iteratively classify images of western art from broad to more specific features, or sequentially testing for presence (or absence) across a range of types of features.

Use Case 2 describes a more bottom-up approach to developing capabilities to produce machine-generated captions for works of art. In this case, training a performant model requires images to be manually annotated. This approach relies either on experts or novices to generate captions, resulting in a folksonomy (an uncontrolled vocabulary) unless a predefined scheme of themes and features is enforced. An example of such a crowdsourcing effort is the steve.museum project [16]. This project provided a web-based application to enable user-generated descriptions of art works in four U.S. museums. These descriptions were then made available to search for and access images.

Use Case 2 discusses some novel ideas to simplify and improve accuracy of image recognition when processing cultural materials. Rather than starting from scratch, the use case leverages the large body of existing descriptions already linked to images to generate new visual descriptions. For example, annotating features with textual labels, rather than the whole work, to improve accuracy in recognition of metaphoric or symbolic imagery. Another tactic employed is identifying relevant sentences that describe the content of a painting within museum catalogs, scholarly articles, or books, and associating them with a specific work of art, or feature in an artwork. Both methods exercise natural language processing (NLP) and named entity extraction to improve the accuracy of automated image annotation.

Use Case 3 focuses on automating the subject description of research datasets, but the problem and the solutions are applicable to all types of datasets. One problem is the asymmetric quality of automated results where some categories have plenty of instances and some have too few, resulting in some automated results that are good and others that are not. A general guide for assessing quality is to strive for 80/20, that is, 80% accuracy is considered reliable enough for general applications (except for domains such as medicine where errors are not acceptable). Another solution is to provide a quality or trustworthiness indicator on poorly performant metadata, or on results that include poorly categorized material. This solution is applied in crowdsourced environments such as Wikipedia.

Use Case 3 also makes a very important observation that training models need to be updated to improve their quality over time. With AI, there is too much focus on ML and less on how to create more adaptive machine models. This is a difficult challenge and requires more emphasis and work.

Corrected Proof

4. SUMMARY

We summarize some challenges and possible solutions on automated metadata annotation in Table 2, and discuss each challenge below:

There are not enough examples with which to train categorizers. While there may not be enough examples, there are certainly some. Use case 1 mentions the *Index of Medieval Art* at Princeton which has cataloged some 200,000 images. There may be plenty of well categorized examples, but aggregating them as a shared resource and making them useful as a collection of training datasets for either traditional ML or advanced deep learning [17] remains a large challenge.

Interpreting metaphors and symbols is too difficult for technology. There is no shortage of stories about the failures of image recognition algorithms, whether mistaking a muffin for a puppy, or St. George riding a bicycle instead of a horse. The problem is not simply to accurately identify objects in works of art, but importantly to tell stories and transfer meanings of an image within the culture and historical context in which the image was created. Both Use Case 1 and 2 suggest that doing image processing *in conjunction with* catalogs of metaphoric and symbolic representations such as ICONCLASS would be productive.

Automated indexing may have inconsistent quality across categories. This is an imperfect world. Sometimes there are lots of high-quality examples readily available to train categorizers, and sometimes there are not enough or poor-quality examples. Categories themselves may be more or less difficult to automate depending on their scope, for example, how discrete the category is from other similar or related categories. Methods are needed to indicate the quality or trustworthiness of automated categorization, and processes to routinely update training models and re-process content are needed.

Table 2. Summary of some automated metadata annotation challenges and solutions.

Challenge	Solution
Rich examples with which to train categorizers	Leverage collections of well-categorized images and related rich text.
Metaphoric and symbolic representations	Use enumerated classifications such as ICONCLASS to recognize simple objects and infer deeper meaning.
Asymmetric quality of results	Provide quality or trustworthiness indicators on automated indexing results.

AUTHOR CONTRIBUTIONS AND ACKNOWLEDGEMENTS

Joseph Busch (jbusch@taxonomystrategies.com) organized a panel "Why AI ≠ Automated Indexing: What Is and Is Not Possible" that presented an early version of these materials at DCMI Virtual 2021. Mingfang Wu (mingfang.wu@ardc.edu.au) suggested getting the group to produce a jointly authored paper for this Special Issue of Data Intelligence. Joseph Busch, the contact author, and Mingfang Wu coordinated this effort. Mingfang Wu and Margorie Hlava (mhlava@accessinn.com) authored the Introduction. Hans Brandhorst (jpjbrand@xs4all.nl) authored Use Case 1 about how to build AI that can recognize iconography and metaphor. Quim Lopez (joaquim.morelopez@bsc.es) and Maria-Cristina Marinescu (mariacristina.marinescu@gmail.com) authored Use Case 2 based on the Europeana St. George on a Bike project that

Corrected Proof

aims to build AI that can recognize "culture, symbols, and traditions." Mingfang authored Use Case 3 about automated metadata annotation of datasets. Joseph Busch authored the discussion and summary sections of the paper.

Thanks to Osma Suominen (National Library of Finland) for reviewing the paper and making suggestions on how to improve it, and to Maureen McClarnon (Taxonomy Strategies) for copyediting the paper.

REFERENCE

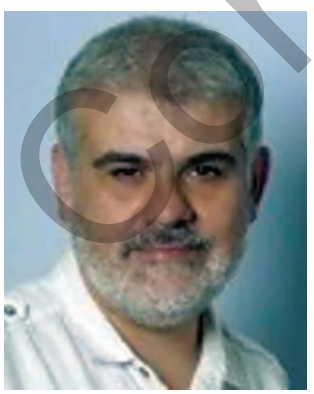
[1] Teztecch. Artificial Intelligence: AI. September 24, 2021. https://medium.com/@teztecch/artificial-intelligence-is-the-simulation-of-human-intelligence-processes-by-machines-especially-5ca1b0c828ec.
[2] Yampolskiy, R., Fox, J.: Artificial general intelligence and the human mental model. In Singularity Hypothesis: A scientific and philosophical assessment, edited by Ammon H. Eden, James H. Moor, Johnny H. Speaker, and Erik Steinhart, Heidelberg: Springer, pp. 129–45 (2012)
[3] Riley, J.: Understanding Metadata: What is metadata, and what is in it for: A Primer. NISO, January 01, 2017. ISBN: 978-1-937522-72-8
[4] Machine Learning, Libraries, and Cross-Disciplinary Research: Possibilities and Provocations. Hesburgh Libraries, University of Notre Dame. https://doi.org/10.7274/r0-wxg0-pe06 (2020)
[5] Maringanti, H., Samarakoon, D., Zhu, B.: Machine learning meets library archives: Image analysis to generate descriptive metadata. LYRASIS Research Publications. https://doi.org/10.48609/pt6w-p810 (2019)
[6] Suominen, O., Inkinen, J., Lehtinen, M.: Annif and Finto AI: Developing and implementing automated subject indexing. JLIS.It 13(1), 265–282. https://doi.org/10.4403/jlis.it-12740 (2022)
[7] Cai, H., Wu, Q., Corradi, T., Hall, P.: The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs. arXiv:1505.00110. http://arxiv.org/abs/1505.00110 (2015)
[8] Crowley, E.J., Zisserman, A.: The art of detection. In European Conference on Computer Vision. Springer, pp. 721–737 (2016)
[9] Milani, F., Fraternali, P.: A dataset and a convolutional model for iconography classification in paintings. Journal on Computing and Cultural Heritage 14(4), Article 46, pp. 1–18. https://doi.org/10.1145/3458885 (2021)
[10] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.F.: ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. doi: 10.1109/CVPR.2009.5206848 (2009)
[11] Cetinic, E., Lipic, T., Grgic, S.: A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE Access 7, pp. 73694–73710 (2019)
[12] Borgman, C.L.: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press (2015)
[13] Wilkinson, M., Dumontier, M., Aalbersberg, I., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18 (2016)
[15] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification Algorithms: A survey. Information 10(4), 50 (2019)
[16] Trant, J.: Social classification and folksonomy in art museums: Early data from the steve. Museum Tagger Prototype. A paper for the ASIST-CR Social Classification Workshop, November 4, 2006. https://www.archimuse.com/papers/asist-CR-steve-0611.pdf
[17] Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? Int J Comput Vis 119, pp. 76–92. https://doi.org/10.1007/s11263-015-0812-2 (2016)
[18] Wu, M., Liu, Y.-H., Brownlee, R., Zhang, X.: Evaluating utility and automatic classification of subject metadata from research data australia. Knowledge Organization 48(3), 219–230. https://doi.org/10.5771/0943-7444-2021-3-219 (2021)

Corrected Proof

AUTHOR BIOGRAPHY

Hans Brandhorst is an independent art historian, editor of the Iconclass system and Arkyves. Together with Etienne Posthumus, he has created the online Iconclass browser and the Arkyves website. He has published on illuminated manuscripts, emblems and devices, iconography and classification, and digital humanities. He was trained in art history at Leiden University and has been using Iconclass as an iconographer since the 1980s. His primary research focus is the simple question "What am I looking at?" in an iconographical sense. His theoretical work deals with the issues of how humanities scholars, in particular iconographers, can collaborate and enrich each other's research results rather than repeat and duplicate efforts. Recently he founded, together with André van de Waal, the "Henri van de Waal Foundation", dedicated to iconographic research with the help of modern technology such as artificial intelligence and machine learning.
ORCID: 0000-0001-8403-3552

Joseph Busch is the founder and principal analyst of Taxonomy Strategies, an organization that works with global companies, government agencies, and NGO's in developing and implementing metadata frameworks and taxonomy strategies. He has extensive knowledge and experience developing content architectures consisting of metadata frameworks, taxonomies and other information retrieval methods to implement effective information management applications such as search engines, portals, websites, content management systems, digital asset management systems, document management systems, knowledge management systems, e-learning, and e-government.
ORCID: 0000-0003-4775-8225

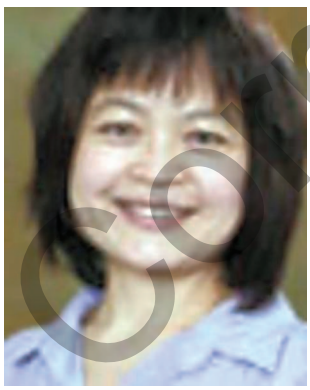


Dr. Joaquim Moré López is a senior researcher and expert in Computational Linguistics at the Barcelona Supercomputing Center (BSC). He has a Ph.D. in Knowledge and Information Society at the Open University of Catalonia. His major areas of expertise are Machine Translation, Information Extraction, Text Mining, Natural Language Processing, Knowledge Engineering, and Opinion Mining. He is working actively in the BSC for the Oficina Técnica de Gestión for the Plan Nacional de Impulso a las Tecnologías del Lenguaje, sponsored by the Spanish Ministerio de Asuntos Económicos y Transformación Digital, to use HPC to exploit the possibilities of Natural Language Processing for Public and Private institutions. He provides solutions to issues related to natural language processing in the Saint George on a Bike project.
ORCID: 0000-0001-5432-0657

Corrected Proof

Marjorie Hlava is the President of Access Innovations, a pioneer at creating explainable AI, she developed the Data Harmony software suite to increase search accuracy and consistency while streamlining the clerical aspects in editorial and indexing tasks. Her most recent innovation is applying those systems to medical records for medical claims compliance in a new application called Access Integrity. She is the author of multiple books and more than 200 articles. She holds two U.S. patents encompassing 21 patent claims.
ORCID: 0000-0002-1624-7695

Dr. Maria-Cristina Marinescu is a senior researcher at the Barcelona Supercomputing Center where she is working on the Smart and Resilient Cities research program. She has a Ph.D. in Computer Science from the University of California, Santa Barbara.
ORCID: 0000-0002-6978-2974



Dr. Mingfang Wu is senior research data specialist at the Australian Research Data Commons (ARDC). She has conducted research in the areas of interactive information retrieval, search log analysis, interfaces supporting exploratory search and enterprise search. Her recent research focuses on the data discovery paradigms as part of the Research Data Alliance (RDA) initiative and for improving data discovery service of Australian national research data catalogue, as well as a few data management related topics such as data provenance, data versioning and data quality.
ORCID: 0000-0003-1206-3431

Corrected Proof