A Comparative Approach between Different Computer Vision Tools, Including Commercial and Open-source, for Improving Cultural Image Access and Analysis

José Luis Preza Díaz
Austrian Centre for Digital Humanities and Cultural Heritage
Austrian Academy of Sciences
Vienna Austria
JoseLuis.PrezaDiaz@oeaw.ac.at

Amelie Dorn
Austrian Centre for Digital Humanities and Cultural Heritage
Austrian Academy of Sciences
Vienna, Austria
Amelie.Dorn@oeaw.ac.at

Gerda Koch
AIT Forschungsgesellschaft mbH
Europeana Local AT
Graz, Austria
kochg@europeana-local.at

Yalemisew Abgaz
Adapt Centre
Dublin City University
Dublin, Irland
Yalemisew.Abgaz@adaptcentre.ie

***Abstract*— Digital cultural heritage objects can benefit greatly from the application of Artificial Intelligence such as computer vision based tools to automatically extract valuable information from them. Novel methods and technologies have been used in the last few years to perform image classification, object detection, caption generation, and other techniques on different types of digital objects from different disciplines. In this pilot study, carried out in the context of the Digital Humanities project ChIA, we present an approach for testing different commercial (Clarifai, IBM Watson, Microsoft Cognitive Services, Google Cloud Vision) and open-source (YOLO) computer vision (CV) tools on a set of selected cultural food images from the Europeana collection with regard to producing relevant concepts. The project generally aims at improving access to implicit cultural knowledge contained in images, and increase analysis possibilities for scientific research as well as for content providers and educational purposes. Preliminary results showed that not only quantitative output results are important, but also the quality of concepts generated. Types of digital objects can pose a challenge to CV solutions.**

Keywords—Artificial Intelligence, Computer Vision, image analysis, cultural heritage

I. Introduction

In recent years, Artificial Intelligence (AI) has attracted much attention across different disciplines, and has seen wide ranging development and application not only in scientific disciplines but also among Cultural Heritage institutions or cultural content aggregators [1]. Among the more recent fields of application is the GLAM (Galleries, Libraries, Archives and Museums) sector, which has drawn on AI applications for opening up access to archives or providing increased user experiences and engagement [2,3]. In the context of cross-disciplinary research, AI tools such as Computer Vision (CV), which can be defined as "the construction of explicit, meaningful descriptions of physical objects from images" (p.xiii,[4]), have frequently been employed to facilitate cultural heritage image access and preservation [5] and database management [6]. As far as the processing of digital objects or images is concerned, a number of novel methods and technologies have been used in the last few years to perform image classification [6], object detection, caption generation, and other techniques on different types of digital objects across different disciplines [7-9]. Extensive research has been done to facilitate content-based image retrieval [10] but the high variability of content and environmental parameters in the cultural heritage domain make the problem quite complex and a precise delimitation of the object of investigation is an important first step [11]. This work focuses on digital cultural heritage data with a particular focus on the aspect of "food", testing different types of digitized images such as paintings, photos or drawings, in black and white or colour. The data, provided by Europeana (https://www.europeana.eu), has previously been curated and provided to Europeana by cultural organisations, including museums and galleries. In particular, we aim to test Computer Vision on a set of selected Europeana cultural food images with the aim to corroborate whether the application of such AI tools can enhance the access to implicitly contained cultural knowledge, contributing to increased knowledge access and analysis possibilities. Our study reports findings from a pilot experiment, which aims to test different commercial and open source Computer Vision systems, specifically comparing the "concepts" predicted by each tool, and focus solely on "object detection", which is the ability to identify objects (with or without bounding boxes) present on an image. In particular, we are interested to find out and report, in how far CV tools are effective in supporting the enrichment of images with additional information to all the generation of new insights and connected knowledge. CV is a wide field with areas of work including image classification, object detection or semantic segmentation, and this study addresses specifically the area of object detection, including object tagging.

The paper is realised in the context of the current Digital Humanities project ChIA (*accessing and analysing cultural images with new technologies*) (https://chia.acdh.oeaw.ac.at/). ChIA aims at testing established state-of-the-art tools (i.e. semantic tools) as well as less typically applied computational methods (Computer Vision) on a selected set of Europeana cultural food images in order to improve image access and analysis, and provide better ways for content providers and educational purposes to gain and make use of to-date unaccessed knowledge. Our approach differs from other studies, in that it focuses solely on cultural heritage objects related to food. It makes a valuable contribution to the field of

978-1-7281-6760-2/20/$31.00 ©2020 IEEE

Digital Humanities as well as to the use and re-use of digital cultural heritage. Cultural data collections can benefit greatly from semantic and CV analysis, particularly with regard to a) metadata validation, where CV results can help to validate existing manual metadata; b) metadata augmentation, where CV can assist with the automatic generation of additional metadata to further enrich a collection, making it easier for searching, segmenting and analysing the data; and c) supporting data according to the FAIR Principles [12], where CV results can have a direct impact on the findability of the data, aligning it more closely to the fact that "data should be findable, accessible, interoperable and reusable" [13].

II. Data & Methodology

A. The Data

For this pilot study, a total of fifteen cultural food images were selected from the online Europeana image collection (~60 million digital objects across different categories and copyright licence types), with "Free to Use" licence only. The data includes a sample of different types of digital images: paintings, photos, drawings, both in black and white or colour. The fifteen images were chosen according to the three selected categories, relevant for the analyses which were a) paintings, b) drawings and c) photographs, and there were five images for each category (see Fig. 1). Specifically, we are dealing with images depicting food items, in a particular "cultural" setting involving either persons, locations, or objects, or a combination of these.

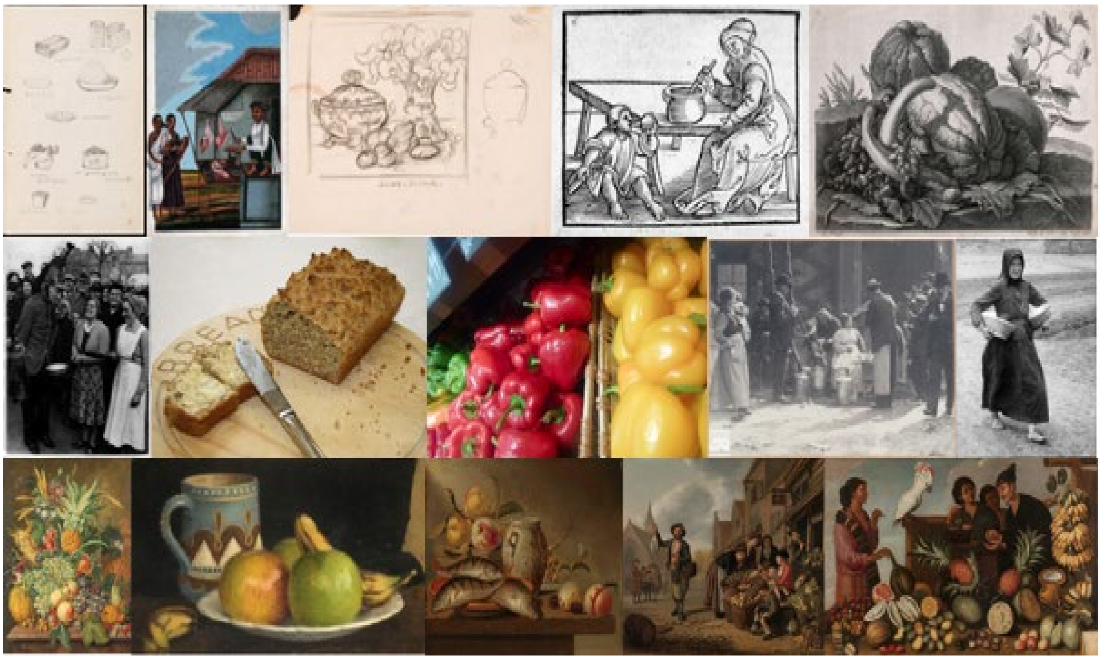

Fig. 1. Examples of selected Europeana cultural food images according to the categories drawings (top panel), photographs (mid panel) and drawings (lower panel).

B. Processing Methodology

After the image selection process, each image was processed with five different Computer Vision (CV) tools with pre-trained models, of which four were commercial tools (Clarifai: general model [14] and food model [15]; IBM Watson: general and food model [16]; Microsoft Cognitive Services: general model [17] (only one available); Google Cloud Vision: general model [18] (only one available) and one open source solution, YOLO [19]. The CV tools were selected based on their level of success in the marketplace [20] and availability. For all commercial tools, the latest online version (mid January, 2020) was used. For YOLO the latest version (version 3) was used. YOLO with the pre-trained weights of the OpenImages dataset, was used with the weights from the pre-trained coco dataset, Darknet, Tensorflow and Keras. The python software code to perform the predictions was forked from [21] and each image was processed with YOLO using this code. For some solutions, a general and a specific food model were available. In this case the general model was applied first, followed by the food model to provide more granular information. Each image was processed using the online version of their solution. The results were then copied to spreadsheets for formatting and calculations. Each individual solution provided similar, but differently named outputs. To enable a comparison between the results of the different solutions, the naming convention and format were standardized. We use the terms "concept" for the class, category or label predicted by a given tool. Similarly, we use the term "probability" for the score, value or statistical probability of the predicted concept. As a next step, "wrong" concepts were identified manually for each image by two evaluators after careful visual inspection of each image, and Cohen's Kappa [22] calculated to determine interrater reliability. By wrong concepts we refer to those concepts returned by a given tool per image, that is actually not depicted, or not plausible in the image context.

III. Results

This section presents the results of the CV analysis for each of the three categories, paintings, drawings and photographs separately. For each category results of the five images were pooled and averages calculated for the total number of concepts. First, we present a numeric overview of the concepts generated by all tools across categories (Table 1). Table 1 and Figure 1 present the CV output for all fifteen images processed, for each of the three categories, paintings, drawings and photographs respectively.

TABLE I. SUMMARY OF THE TOTAL NUMBER OF OVERALL CONCEPTS, POOLED ACROSS ALL PAINTINGS, PHOTOGRAPHS AND DRAWINGS (N=15)

Tool/Category	*Photos*	*Paintings*	*Drawings*	*TOTAL*
YOLO-coco	22	14	3	**39**
Microsoft	111	73	46	**230**
Google Vision	114	109	49	**272**
Claraifai Gen.Model	100	100	100	**300**
Clarifai Food Model	31	95	12	**138**
IBM Watson General Model	52	50	48	**150**
IBM Watson Food Model	14	8	5	**27**
TOTAL	**444**	**449**	**263**	**1156**

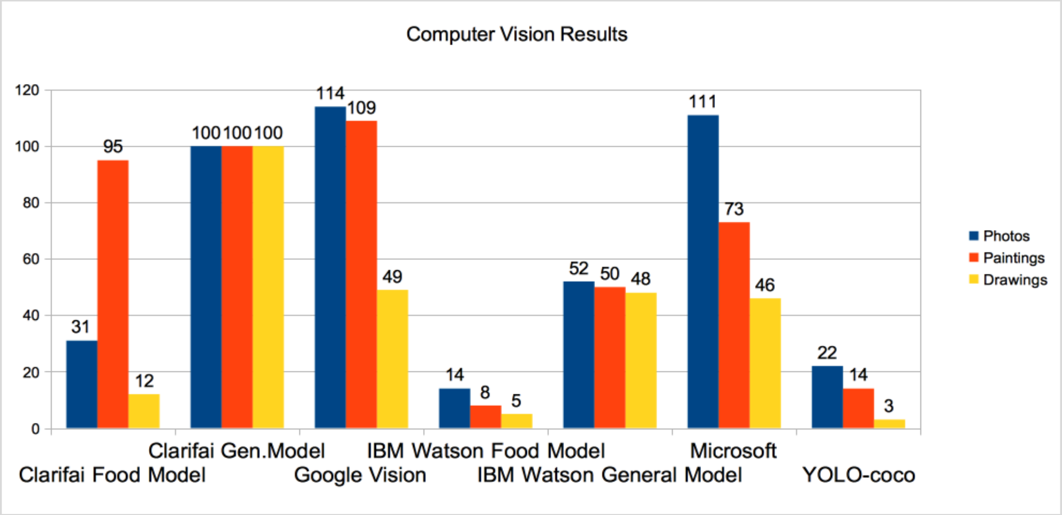

Fig. 2. Overview of all concepts generated per tool across the three categories photos (blue bars), paintings (orange bars) and drawings (yellow bars).

Overall, we can state that there is considerable variation between the numbers of concepts generated by each CV tool, as well as across the three categories. In total 1156 concepts were generated. The highest number of concepts per category was generated for paintings (n=449), followed by photos (n=444) and drawings (n=263). In terms of CV tools, the highest number of concepts across the three categories was generated by the Clarifai general model (n=300), and the smallest number by the Watson food model (n=27).

To rate accuracy of the concepts identified as right or wrong between the two evaluators, the inter-rater method was used and Cohen's Kappa coefficient (κ) calculated per image, then averaged across all objects for each of the three categories separately (see Tables II, III and IV).

TABLE II. KAPPA VALUES FOR EACH PAINTING (P) AND AVERAGES (AVG) FOR ALL IMAGES PROCESSED (N=5)

Tool / Object	*P1*	*P2*	*P3*	*P4*	*P5*	*AVG*
YOLO-coco	1	1	1	1	1	**1**
Microsoft	0.66	1	1	1	1	**0.93**
Google Vision	0.42	0.47	0.46	0.76	0.64	**0.55**
Clarifai Gen.Model	-0.17	0.77	0.89	0.86	0.77	**0.63**
Clarifai Food Model	0.77	0.87	0.61	0.87	1	**0.63**
IBM Watson Gen.Model	0.56	1	1	1	1	**0.93**
IBM Watson Food Model	1	1	1	1	1	**1**

Looking first at the category "paintings" (Table II), we can observe that perfect agreement was reached for YOLO and IBM Watson food model (κ=1). Only moderate agreement was reached on concepts generated by Google Vision (κ=0.56).

Looking next at the category "photographs" (Table III), we observe again that perfect agreement was again reached for concepts generated by YOLO and the IBM Watson food model (κ=1). Substantial agreement was found for concepts generated by the Clarifai general model (κ=0.73). Note for the Clarifai food model, that three out of the five images could not be processed.

TABLE III. KAPPA VALUES FOR EACH DRAWING (D) AND AVERAGES (AVG) FOR ALL IMAGES PROCESSED (N=5)

Tool / Object	*D1*	*D2*	*D3*	*D4*	*D5*	*AVG*
YOLO-coco	n/r	n/r	n/r	n/r	1	**1**
Microsoft	0	1	1	0.57	1	**0.71**
Google Vision	0	0.63	1	1	1	**0.73**
Clarifai Gen.Model	0.61	0.77	0.88	0.69	0.61	**0.71**
Clarifai Food Model	n/p	0.43	n/p	n/p	n/p	**0.43**
IBM Watson Gen. Model	0.53	0.83	0.55	0.14	0.76	**0.56**
IBM Watson Food Model	1	1	1	1	1	**1**

Looking finally at the category "drawings" (Table IV), we observe also for this category that perfect agreement was again reached for concepts generated by YOLO and the IBM Watson food model (κ=1). Only moderate agreement, however, was found for the Clarifai food model (κ=0.43), where four out of the 5 images could not be processed with this tool. Similarly, YOLO did not yield any results for four out of the 5 images.

TABLE IV. KAPPA VALUES FOR EACH PHOTO (F) AND AVERAGES (AVG) FOR ALL IMAGES PROCESSED (N=5)

Tool / Object	*F1*	*F2*	*F3*	*F4*	*F5*	*AVG*
YOLO-coco	1	1	1	1	1	**1**
Microsoft	0.43	1	0.75	1	0.84	**0.81**
Google Vision	1	1	0.76	1	1	**0.95**
Clarifai Gen.Model	0.69	0.5	1	0.46	1	**0.73**
Clarifai Food Model	N/P	N/P	N/P	0.74	1	**0.87**
IBM Watson Gen. Model	0.71	0.84	1	0.62	1	**0.83**
IBM Watson Food Model	1	1	1	1	1	**1**

In addition, we also determined a numerical overview of different food related concepts generated across the three categories. By filtering the concepts obtained from the processing by the terms that include "food" in them, it is possible to discover which objects have been associated to "food" by the different systems. In total, 15 different concepts found by the systems, all related to "food", were observed in a total of 39 occurrences. Figure 3 shows a network graph generated with Apache Superset [23], showing links between the food-related concepts and image categories. Additional work is needed to filter the data by the correct concepts.

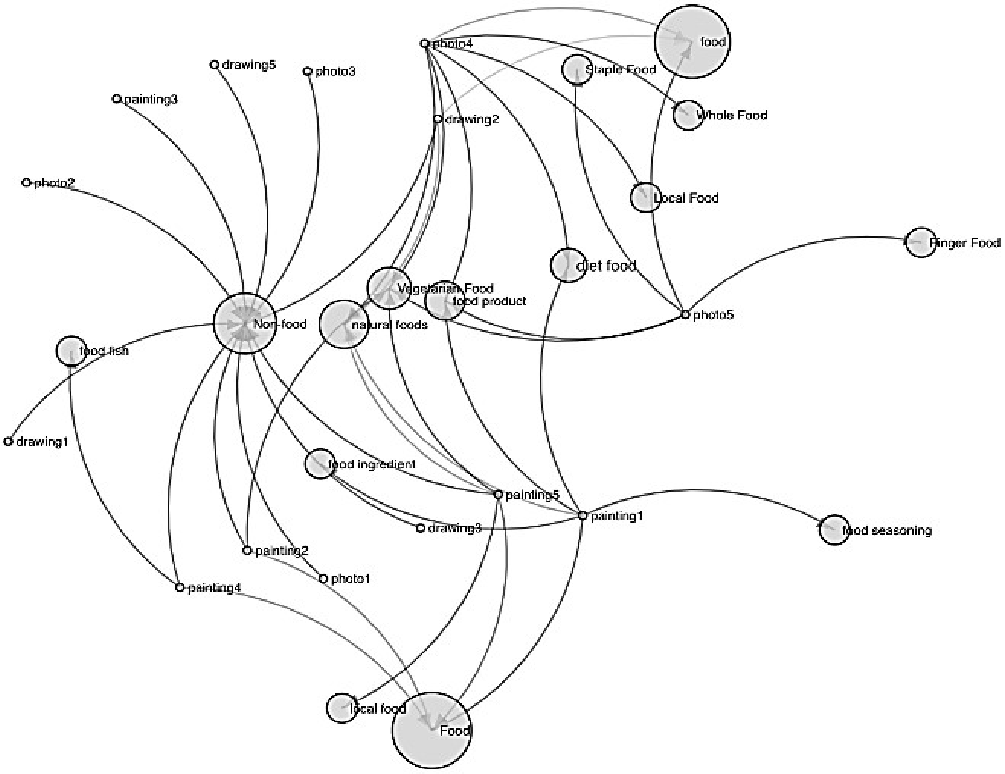

Fig. 3. Superset network graph of all generated concepts per category of objects related to "food". Bubble size indicates higher number of occurrences of the same concept.

IV. DISCUSSION

This pilot study tested a number of commercial and open-source CV tools on a selected set of cultural food images across three categories (paintings, drawings, photographs) for comparing their performed output in terms of generated concepts, for the purpose of enriching the images with additional, relevant food-related and cultural data. Overall we can summaries that not only the quantity, but also the quality of outputs generated by CV solutions are important for successfully enriching cultural food images. As to quantitative outputs, we have observed that there is considerable variation between the absolute numbers of concepts generated per CV tool, as well as across categories. In addition, not all image categories could equally successfully be processed by each of the tested tools. Drawings, in particular sketches, posed a challenge to any of the CV solutions tested, given their often unfinished and less detailed nature. In addition, to fully measure accuracy, it is also important to take into account the total number of different concepts predicted as well as the number of images where concepts were predicted at all. For example, there were several objects where YOLO did not predict any concept. But not only the amount of concepts predicted is important, but also their quality. Another important aspect is the granularity of the concepts predicted. Ideally both, detailed and higher level concepts (e.g., banana for detail, food for higher level) should be obtained. Obtaining only higher levels concepts might not be the best option as it is not possible to go down from a higher level to a more granular one, but the opposite is possible when processing the detailed concepts with a lexical tool like Wordnet. With solutions like Clarifai though this is not an issue. If with this tool anything related to "food" is detected on an image when processed with the general model, the concept "food" will be predicted, this indicates the need to run the image using the food model which in turn will identify the detailed information regarding "food". Another concrete example is the concept "person". YOLO returns such concepts, not identifying if man or female. This, however, is not automatically wrong, it depends on the need and the application. Another pertinent aspect that emerged during the analysis concerns the marking of concepts as wrong or right. The identification of wrong concepts is a cognitive process that has to be undertaken by a human as the validation of predicted concepts via automatic processes (validated against manual metadata, for example), although technically possible, might not be optimal. This is mainly due to the fact that the manual metadata does not contain all information that computer vision algorithms predict. And it not only deals with identifying totally wrong concepts, but also concepts that might or might not be wrong. For example the result "juice" for painting1, where the image depicts a mug. The mug could contain a liquid, it could also be empty, neither a person nor an algorithm could possibly identify its contents from seeing the mug as depicted. In this case, “juice” might not be totally “wrong”. Considering key factors including the average number of concepts returned and the type of information capable of producing, for the purposes of ChIA, Clarifai offered the most suitable outputs, followed by Google Cloud Vision, Microsoft, IBM Watson and YOLO. Something that also proved beneficial for the analysis of cultural food images, was the application of both the general followed by the food models. This method provided more granular information on food objects, not only detecting food in general, but also giving information on ingredients. A similar method could also be employed for other details such as colour, texture, demographics, etc. Overall we can summarise that it is not enough that a platform delivers concepts that are accurate, but the amount of concepts delivered as well as their probability are important factors to the equation.

V. Conclusion & Outlook

Concluding we can state that all systems tested in this pilot experiment delivered valuable information, which can add to the increased access analysis possibilities for digital images. The results from each tool are in most cases complementary to each other, and for each application of CV solutions it is relevant to consider the desired results and purpose of the application. Depending upon this, a single result might be desired, in other cases it would be useful to recognize as many objects in an image as possible, thus making use of several solutions simultaneously and aggregating their results could prove the most satisfactory solution. As this study only offered a first testing of solutions given the early stage of the project, we intend to extend the number of parameters (e.g. performance benchmarks; colour or demographic information), potentially also including technical metadata [24,25]. As the output is currently flat, putting results into a structured hierarchy, using a classification method (e.g. WordNet), could prove beneficial to organise results in a more structured way, allowing to cluster and search data in higher definitions. Finally we can conclude that CV offers potential for increasing the access and analysis methods for cultural food images, but more data will need to be processed for more conclusive results.

Acknowledgment

This research is funded by the Austrian Academy of Sciences under the funding scheme: go!digital Next Generation (GDNG 2018-051). The ChIA project is carried out in collaboration with the Adapt Centre, DCU. ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund."

References

[1] Bordoni, Luciana, Mele, Francesco, Sorgente, Antonio (2016). Artificial Intelligence for Cultural Heritage. Cambridge Scholars Publishing.

[2] Simon, Nina (2010).The participatory Museum. Museum 2.0.

[3] Styx, Lauren. (2019).How are museums using artificial intelligence, and is AI the future of museums? Museum Next. Online article.
https://www.museumnext.com/article/artificial-intelligence-and-the-future-of-museums/ [last accessed: 02.02.2020]

[4] Ballard, D.H. & Brown, C. M.(1982). Computer Vision. Prentice Hall. First edition.

[5] Hardman, L., Van Ossenbruggen, J., Aroyo, L., & Hyvönen, E. (2009).Using AI to Access and Experience Cultural Heritage. Intelligent Systems, IEEE 24(2):23 - 25.

[6] Chia-Ching Hung, (2018) "A study on a content-based image retrieval technique for Chinese paintings", The Electronic Library, Vol. 36 Issue: 1, pp.172-188, https://doi.org/10.1108/

[7] Lo Brutto, Mauro & Meli, Paola. (2012). Computer Vision Tools for 3D Modeling in Archaeology. International Journal of Heritage in the Digital Era. 1. 1-6. 10.1260/2047-4970.1.0.1.

[8] Behli, Abdelhak et al. (2018). Leveraging Known Data for Missing Label Prediction in Cultural Heritage Context. https://doi.org/10.3390/app8101768

[9] Ciecko, Brendan. “Examining the Impact of Artificial Intelligence in Museums.” MW17: MW 2017. Published February 1, 2017. Consulted March 26, 2017. http://mw17.mwconf.org/paper/exploring-artificial-intelligence-in-museums/

[10] Zujovic, Jana & Gandy, Lisa & Friedman, Scott & Pardo, Bryan & Pappas, Thrasyvoulos. (2009). Classifying paintings by artistic genre: An analysis of features & classifiers. 2009 IEEE International Workshop on Multimedia Signal Processing, MMSP '09. 1 - 5. 10.1109/MMSP.2009.5293271.

[11] Tzouveli,P., Simou, N.,Stamou, G. & Kollias, S. (2009). Semantic Classification of Byzantine Icons. IEEE Intelligent Systems (Volume: 24, Issue: 2, March-April 2009)

[12] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016).The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[13] Preza, J.L. (2016). Automated Information Enrichment for a Better Search. Report. Zenodo. http://doi.org/10.5281/zenodo.163933

[14] Clarifai General Model. https://www.clarifai.com/models/general-image-recognition-model-aaa03c23b3724a16a56b629203edc62c [last accessed 02.02.2020]

[15] Clarifai Food Model. https://www.clarifai.com/models/food-image-recognition-model-bd367be194cf45149e75f01d59f77ba7 [last accessed 02.02.2020]

[16] IBM Watson. General and food model. https://watson-visual-recognition-duo-dev.ng.bluemix.net/pre-trained [last accessed 02.02.2020]

[17] Microsoft Cognitive Services. General model. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/[last accessed 02.02.2020]

[18] Google Cloud Vision. https://cloud.google.com/vision/[last accessed 02.02.2020]

[19] YOLO – Real-Time Object Detection. Redmon, J. & Farhadi, A. (2018) YOLOv3. https://pjreddie.com/darknet/yolo/ [last accessed 02.02.2020]

[20] Forrester Report, Computer Vision. The Forrester New Wave Online article. https://cloud.google.com/forrester-computer-vision/ [last accessed: 01.02.2020]

[21] https://github.com/experiencor/keras-yolo3 [last accessed:01.02.2020]

[22] Cohen J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement.1960;20(1):37–46.

[23] Apache Superset. https://superset.incubator.apache.org/

[24] Zartl, A. (2016). Automatische Übernahme von technischen Metadaten https://phaidra.univie.ac.at/view/o:462586

[25] Blumesberger, S. (2018). Metadaten als Mehrwerte: Konvergente Strategien, Methoden und Konzepte. 10.1515/9783110539011-018.