Journal of Open Humanities Data

The EyCon Dataset: A Visual Corpus of Early Conflict Photography

DATA PAPER

ubiquity press

MARINA GIARDINETTI
DANIEL FOLIARD
JULIEN SCHUH
MOHAMED-SALIM AISSI

*Author affiliations can be found in the back matter of this article

ABSTRACT

The EyCon dataset, comprising nearly 130,000 JPEG images and pages, documents armed conflicts from the 1890s to 1918, with a focus on extra-European contexts. The project team aggregated thousands of digitized images and metadata from various institutions, including previously inaccessible documents. To enhance metadata, the team conducted visual and multimodal similarity analyses, as well as human and animal detection. Captions were processed to extract named entities for XML-formatted descriptive metadata. Challenges in identifying and publishing graphic images due to automated tools' limitations in detecting violence were addressed with human expertise for accurate classification. Available online and on Zenodo for download and reuse, the dataset confronts issues in computer vision for heritage photographs, such as degradation from fading, discoloration, scratches and noise, which impair algorithms reliant on visual features. The under-representation of early photographic cultures in datasets introduces bias in applying standard solutions to archival materials.

CORRESPONDING AUTHOR:
Marina Giardinetti
LARCA, Université Paris-Cité, Paris, France
marina.giardinetti@gmail.com

KEYWORDS:
visual heritage; photography; newspapers; deep learning; multimodal similarity; named entity recognition

TO CITE THIS ARTICLE:
Giardinetti, M., Foliard, D., Schuh, J., & Aissi, M.-S. (2024). The EyCon Dataset: A Visual Corpus of Early Conflict Photography. *Journal of Open Humanities Data*, 10: 40, pp. 1–12. DOI: https://doi.org/10.5334/johd.213

(1) CONTEXT AND MOTIVATION

(1.1) REPOSITORY LOCATION

This dataset is part of an extended report about the outcomes of the Idex Université Paris Cité/AHRC/Labex Passés dans le présent funded project "EyCon – Visual AI and Early Conflict Photography". The dataset is stored in a Zenodo repository (Foliard et al., 2024) with DOI: 10.5281/zenodo.11449122 and accompanied by a list of archival references for the collections used in the datasets as well as additional documentation on data curation. It is also searchable via an online platform at https://eycon.huma-num.fr/s/en/page/accueil. Additional information on the project can be found at https://eycon.hypotheses.org/.

(1.2) CONTEXT

Recent digitization efforts of historical photographs by archival institutions have often been conducted in silos. This presents a significant challenge for researchers and archivists and raises concerns regarding the public use of history, particularly in the context of contemporary perspectives on colonial and imperial warfare. Disconnected visual repositories perpetuate entrenched notions of national exceptionalism in countries such as France, Britain, and other states with histories of international interventionism and expansionism. The EyCon project concentrated on early conflict photography from the period 1890–1918, documenting mass armed violence. This scope included overlooked colonial campaigns and the First World War, particularly focusing on the battlegrounds in Africa and Asia. The primary objective of EyCon was to consolidate fragmented repositories by developing a dataset of digitized photographs that capture lesser-known aspects of modern warfare. The intention was to apply computer vision methodologies to this material (Abgaz et al., 2021), specifically to trace the circulation of photographs from their creation by both amateur and professional photographers to their subsequent use by the press. This initiative aimed to improve the discoverability and usability of overlooked and dispersed materials related to colonial, imperial, and international armed conflicts up to 1918. By doing so, it sought to analyse the visual culture of war and organised violence from the late 19th and early 20th centuries across a range of collections on a transnational scale. The project also aimed to challenge the often limited and Eurocentric perspectives on global warfare, which in France and Britain are heavily influenced by the visual coverage of the two world wars and their Western battlefields. By recovering obscured experiences of conflict, EyCon endeavoured to broaden our understanding of organized violence.

The project team collaborated with a broad consortium of institutional partners in both France and the United Kingdom, countries pivotal in colonial expansion and global interventionism during the late 19th and early 20th centuries. In this endeavour, EyCon facilitated connections among archivists and experts from leading institutions that curate the visual heritage of the so-called "distant" or "small wars" characteristic of peak European colonial expansion in Africa, Oceania, and Asia.

The project applied machine learning and deep learning techniques to analyse 19th and early 20th-century digitized photographic collections. This approach aimed to study the circulation of original prints in the illustrated press, identify similar images held by various institutions (Aske & Giardinetti, 2023), and create a comprehensive database of materials from partner institutions, complemented by semi-automated metadata enrichment. Metadata enrichment of large digitized image collections at the item level has become increasingly important for researchers and archivists (Elo, 2020), with AI tools offering significant assistance in this area (Männistö et al., 2022). Nonetheless, the use of automation in archives documenting contested histories (Foliard, 2020) presents ethical dilemmas (Wevers & Smits, 2020), especially when dealing with data that is entangled in coloniality. Moreover, the relative scarcity of datasets suitable for historical studies containing heritage monochrome photographs and their halftone reproductions complicates obtaining accurate metrics, adding another layer of complexity to structuring such potentially sensitive data.

Historians working within the EyCon framework were able to use the processed data to go beyond conventional historiographical methods that can prove inadequate for fully grasping the extensive circulation, cultural significance, and economic dynamics of these photographs (Schill, 2024). Various research projects have addressed our human limitations when it comes to surveying vast amounts of digitized photographs (Lee et al., 2020). Many walk in the footsteps of L. Manovich and F. Moretti (Manovich, 2020; Moretti, 2013), who both claim that "distant reading" — i.e. the statistical and computerised processing of information from numerous texts without actually reading them — enables us to shift our gaze away from the single archival document, with a view to revealing unseen relationships, circulations, and patterns. Art history writing has traditionally prioritized pictorial analysis, focusing primarily on photographs as individual, aesthetically determined images. However, understanding the complexity of amateur and non-artistic photography from this perspective has proven challenging. To grasp the real complexity and historical variability of the types of images tackled in the project, it is essential to comprehend the systems and structures involved in their creation, circulation, printing, and storage. The period and events under study in the EyCon project were pivotal in establishing 20th-century norms of war coverage and photojournalism. By the early 20th century, the widespread adoption of halftone photomechanical reproduction reinforced the association between photography's supposed indexicality and ideals of objectivity. The project aimed to address the rapid "photo-inflation" of the late 19th and 20th centuries – when millions of images were printed and circulated globally – by focusing on a specific subgroup of images. Their many transformations and trajectories into printed photo engravings and illustrations are hardly considered at scale and across entire sets. Computer Vision and Artificial Intelligence (CVAI) can help detect visual and textual contents, as well as track the circulation of original and similar photographs. CVAI-assisted "distant viewing" (Arnold & Tilton, 2023) can help historians, curators and end-users to augment their capacity to trace the circulation of images, generate image metadata and create datasets targeted at researchers in digital humanities and other fields.

Understanding the intricate systems and mechanisms involved in the creation, dissemination, printing, and archival storage of news photographs is essential to grasping their true complexity and historical diversity. Collecting diverse data aims to assess the dissemination of these photographs, viewing it as a vital aspect of public comprehension. This approach also seeks to mitigate the inaccessibility of historical photographs and their contexts, providing a more comprehensive understanding of their role and impact.

(1.3) DATA COLLECTION

EyCon focused on overlooked extra-European armed conflicts and battlefields from the 1890s to the end of the First World War. Well-known conflicts such as the second Boer War or the so-called Boxer rebellion are documented, as well as less prominent campaigns such as the war against Samory Touré in West Africa. Even if the dataset does not cover all the conflicts of the period exhaustively, it provides a carefully curated selection that reflects Western coverage of extra-European conflict at the turn of the 20th century.

In France, participating institutions included the Service Historique de la Défense in Vincennes and the Établissement de Communication et de Production Audiovisuelle de la Défense in Ivry. Both were established by the French War Office and thus rich in photographic material on French military campaigns worldwide. Other contributors included the Musée du Quai Branly – Jacques Chirac (Paris), the French National Archives (Pierrefitte-sur-Seine), and La Contemporaine (Nanterre), all of which possess collections of original albums and photographic series documenting colonial campaigns. The Archives Nationales d'Outre-mer (Aix-en-Provence), which holds private and official archives from French overseas territories and former dependencies, was also a significant contributor, providing expertise and data, including several rare amateur albums that were digitized within the framework of the EyCon project. UK partners paralleled the French collaborations. The Imperial War Museum (London), with its extensive collections of official and amateur photographs of armed conflict, contributed its expertise and previously untapped data documenting the First World War in Africa. The Wellcome Collection (London) also participated, not only because it curates several relevant amateur and official albums, but also due to its proficiency in managing digital materials. All partners provided both existing data and new material, which the EyCon project aimed to disseminate widely. The project complemented this data with additional images collected from freely available online collections and with newly digitized material. The dataset not only includes already available photographs and prints, but also thousands of previously unavailable materials such as the Album Lartigue, which documents the war waged by France against Samory Touré's empire, or *Sur le Vif*, a French illustrated newspaper created during the First World War.

The dataset encompasses a broad spectrum of photographic objects and techniques. The collection features most of the photographic and printing processes that characterized the turn-of-the-20th-century image culture. It comprises digitized postcards, glass plates, stereoviews, albumen prints, aristotypes, cyanotypes, halftone reproductions from periodicals, rotogravures, and heliogravures. Given that many amateur photographers sold their negatives and prints to newspapers during a period when news image agencies were nascent, the dataset includes unpublished personal albums, individual prints, official and professional sets of photographs, as well as illustrations from books and periodicals. The dataset was curated to include entire sets, thereby avoiding any anachronistic filtration and selection of the archives. For instance, an amateur album might contain a family portrait taken in France alongside views of a battle in Morocco from the 1910s. The images not only include graphic and direct views of combat and its consequences, but also more benign photographs of landscapes, individuals and non-military events.

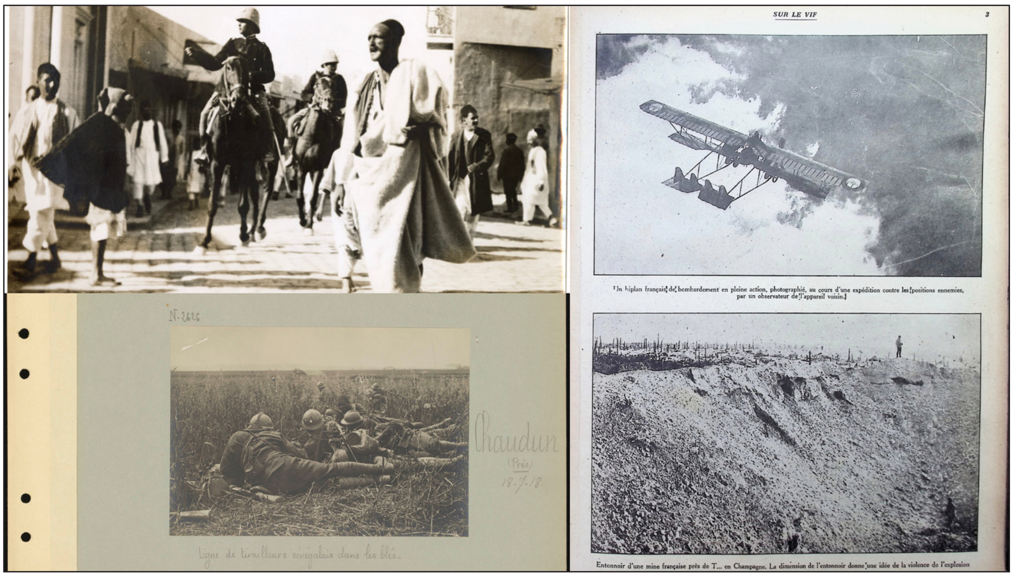

N.2626
Chaudun
(Près)
18-7-18
Ligne de tirailleurs sénégalais dans les blés.
SUR LE VIF
3
Un biplan français de bombardement en pleine action, photographié, au cours d'une expédition contre les positions ennemies, par un observateur de l'appareil voisin.
Entonnoir d'une mine française près de T... en Champagne. La dimension de l'entonnoir donne une idée de la violence de l'explosion

Figure 1 Top left, Lybie – Tripolitaine, 2K47161-70, photographic print, 21,5 × 16,5 cm, Victor Forbin collection, courtesy of Service Historique de la Défense (Vincennes, France); Bottom left, Ligne de tirailleurs sénégalais dans les blés, BDIC-VAL-008-159, Fonds Valois (Aisne), 1918, La Contemporaine (Nanterre, France); Right, *Sur le Vif*, 04/11/1916, page 3 (digitized by the EyCon team).

The dataset comprises over 130,000 documents. It includes entire collections, album and periodical pages and individual photographs in JPEG format. Our partner institutions provided direct access to their documents and metadata. As a result of the differences between the controlled vocabularies of the institutions, new specifications had to be established for search and selection via download APIs (Bibliothèque Nationale de France, Wellcome Collection, Library of Congress). This is particularly obvious for armed conflicts that have been subjected to divergent processes of memorialization and heritagization by nation-states. For instance, Gallica identifies the so-called Boxer War as "Boxers, Guerre des (1895–1900)" (Rameau, n.d.), while the Library of Congress (n.d.) lists it as "China-History-Boxer Rebellion, 1899–1901 -Campaigns & Battles". Defining and indexing documents in our database sometimes required new subject headings, which were discussed in the course of the project. This was specifically true for colonial campaigns in Africa that were not properly indexed by major institutions or Wikidata, a case in point being the absence of any subject heading for the French campaigns against the Wassoulou Empire in the 1880s and 1890s. Our database of digital materials had to be described in novel ways that included rethinking semantics. To ensure alignment and referencing of our data, we aimed to minimize semantic bias by utilizing Wikidata as a bridge between controlled vocabularies and collections. Wikidata enables the identification of people, events, or locations with unique identifiers from various denominations, thereby defining different corpora at the same semantic level. The data collected in the course of the project is the product of a history of domination and imperialism. Much of the visual heritage from the colonial era is curated, both physically and digitally, by institutions in countries that were imperial metropolises in the 19th and 20th centuries. This archival gap is often exacerbated by a broader digital divide, as digitization infrastructures are far more advanced in Europe and the USA than in African countries, whose colonial histories are documented in the photographs addressed by the project.

Two interns digitized additional material, resulting in valuable new collections with precise metadata, in addition to this already existing material. A selection of more than 1,000 original photographic prints from the Rumpf-Forbin collection at the Service Historique de la Défense (Vincennes, France) representing a wide range of conflicts were scanned with high-resolution scanners. Born in 1864 in Paris, Forbin was a journalist, explorer, and writer. His travel stories and articles appeared in the daily press. From the late 1890s onwards, he gradually built up his own news agency, one of the first in France. This role involved providing images from around the world to major French newspapers. 1915 to 1917 copies of the French illustrated paper *Sur le Vif*, which documented the First World War with a rare wealth of photographic illustrations, were also digitized (Figure 1). The team also created a data architecture with several format-based sub-corpora to account for the variety of curated photographic material (individual prints, photographic albums, photographic collections and periodicals and magazines) that were collected within the framework of this project (Figure 2).

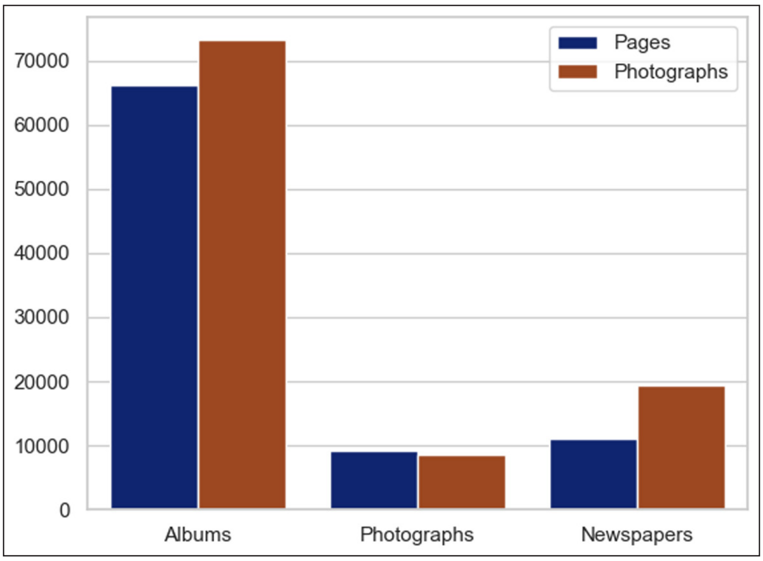


Figure 2 Distribution of images within the three main categories of EyCon documents and the number of pictures extracted.

(2) DATASET DESCRIPTION

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.11449122

REPOSITORY NAME

Zenodo

OBJECT NAME

EyCon project photographs and metadata

FORMAT NAMES AND VERSIONS

JPG – XML

CREATION DATES

2019-12-01 to 2023-12-06

DATASET CREATORS

Daniel Foliard (researcher, LARCA, UMR 8225, Université Paris Cité), Julien Schuh (researcher), Marina Giardinetti (editor, LARCA, UMR 8225, Université Paris Cité), Mohamed-Salim Aissi (LIP6, Sorbonne Université, France), Jonathan Dentler (researcher, Université Paris Nanterre).

LANGUAGE

French and English

LICENSE

Public Domain, various licences apply at item level (cf. https://eycon.huma-num.fr/s/en/page/licences).

PUBLICATION DATE

2024-06-03

(3) METHOD

(3.1) IMAGE AND TEXT EXTRACTION

Image extraction from newspaper and photographic album pages was the first step after data collection from the project partners. We used Layout Parser (Shen et al., 2021) to extract individual items. Two different models were trained on the annotated document pages (Gutehrlé & Atanassova, 2021) via Visual Geometry Group (VGG) Annotator,[1] which allows the creation of data files in Common Objects in Context (COCO) format containing the position of the images: one model for the albums' layout and the other for the newspapers' layouts. Extracted images were added to the dataset.

We also applied object detection to the images. We used a You Only Look Once (YOLO) v3 model pre-trained on recent data (Cifar 100 dataset). Among the pre-trained object classes, human beings and horses were the only two objects chosen for detection because they are more resistant to the historical bias that characterizes existing datasets. Their presence in the pictures was documented by a specific tag in the metadata. The main goal of this extraction was to supplement the original metadata and allow the creation of new research features by visual content.

A visual and multi-modal similarity calculation was also used in order to identify photographs that were similar or had been taken at the same time. Our goal was to extract the most significant subsets from our image collection as well as the largest number of strongly connected subgraphs from our dataset. We applied a pre-trained ImageNet[2] model on our dataset and enhanced it with half toning using the Jarvis algorithm (Aissi, 2023) to compensate for image quality. The use of textual information, such as captions, descriptions, or metadata, allowed us to refine our search and obtain better results. A multimodal approach, which combined visual and textual elements, enabled us to better understand and select results that were historically relevant. The similarity calculation involved a two-step process. First, the visual and textual features of a given data point are combined by concatenating their respective feature vectors. This results in a unified representation that encapsulates both the visual and textual characteristics of the data.

The automated detection and extraction of named entities from captions were performed to index the entire dataset. Characters and locations were identified using a pre-trained SpaCy model,[3] commonly employed for named entity recognition (NER).[4] These identified entities were then aligned with another database, with Wikidata being chosen for its multilingualism and automatic geolocation capabilities.

However, much like visual models trained on ImageNet and COCO, natural language processing (NLP) models provided by SpaCy are characterized by anachronisms and bias (Zhang et al., 2022). While these models facilitate data recognition and localization, they are influenced by linguistic and political biases (Feng et al., 2023) due to the under-representation of non-European languages in NLP tools, which was an issue given the project's perimeter. Furthermore, despite significant ongoing research (Ehrmann et al., 2022), NER for historical archives still faces challenges, including a lack of digitization and the presence of Optical Character Recognition (OCR) noise (Ehrmann et al., 2023).

[1] https://www.robots.ox.ac.uk/~vgg/software/via/.

[2] https://www.image-net.org/.

[3] https://spacy.io/usage/linguistic-features#named-entities.

[4] https://github.com/NER4Archives-project/spacy_ner_training_pipeline.

The project's experiments with automation revealed how critical it is for metadata standards to evolve as AI tools become increasingly prevalent in the humanities and social sciences, particularly in history. To trace the context of information accurately, metadata should account for the origin of each piece of information. When creating the metadata XML files, tags indicating the author of each piece of information are added, which correspond to the pictograms displayed on the website. It is essential to distinguish between AI-generated metadata and metadata created by an archivist, and this distinction should be clearly indicated in the International Image Interoperability Framework (IIIF) manifests. Additionally, accuracy metrics should be included in the metadata. The design of EyCon's online database proposes a method to incorporate this information in a user-friendly manner: metadata is identified by different pictograms to distinguish between automatic and human input. This approach ensures transparency and reliability in the use of AI-reliant metadata in historical research.

(3.2) DATA MODELLING AND STORAGE

Eycon's data structure (Figure 3) was designed to work along the grain of photographic archives and to curate AI-created metadata.

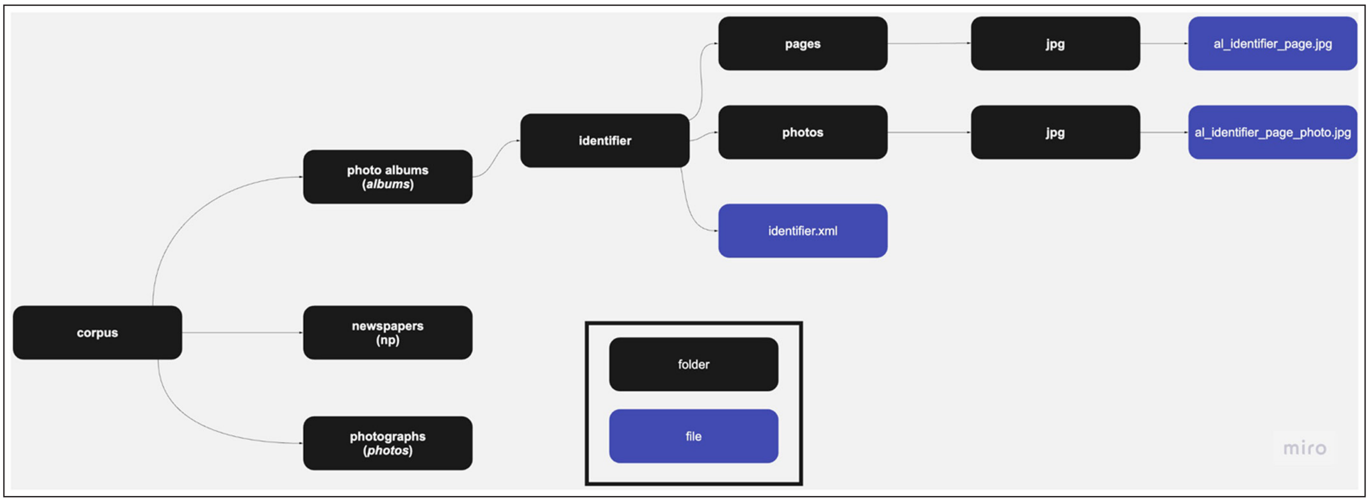


Figure 3 EyCon project's data structure, made with Miro. The stockage is based on the structuration of the material archives.

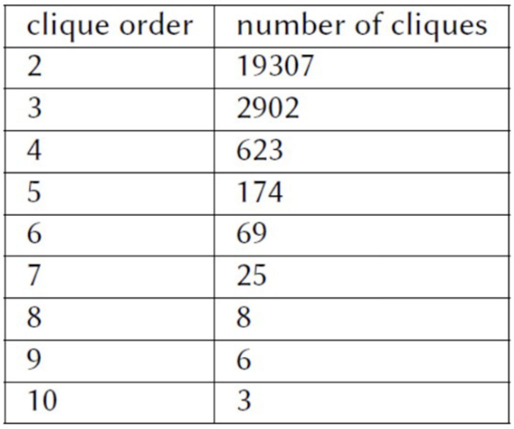
The primary objective in designing the database was to avoid overly detailed descriptions that might overshadow the relationships between images and their contexts of creation. Instead, the focus shifted towards viewing images as more than mere representations, emphasizing their material presence within various contexts such as pages or albums. This approach ensured that attention was not solely on what the photograph represented, but also on its tangible existence and circulation. The specialist in charge of the data model considered both archival formats and automatically generated metadata (NER, object detection, similar images). Data modelling aimed to mirror the granularity of the archives, encompassing entire collections down to individual item-level documents. Designing this structure was crucial to representing the materiality of the digitized archives and avoiding isolating the visual content of a photograph from its context. To facilitate efficient organization and retrieval, a unique identifier was assigned to each photograph based on the type of document, its original classification number, page number, and photograph number (e.g., al_2K194_52_03, with 'al' denoting albums). This identifier system was consistently utilized throughout the preservation process, from physical storage to metadata tagging. Notably, photographs were systematically named, incorporating elements such as document type, unique identifier, date (for newspapers), page number, and photograph number. This meticulous approach ensured that conflicts during data processing were eliminated, guaranteeing clear differentiation and sortable categorization of images, thus preserving the layers of the physical archive.

The descriptive metadata was retrieved in various formats and levels of detail. One of the initial tasks of the data team was to standardize these disparate metadata sets, making them usable by an algorithm and indexable for the database users. To achieve these goals, the various metadata was stored in XML-EAD format, which allows for extensive enrichment. In addition to the metadata supplied by the institutions and converted to XML-EAD format, we augmented each extracted and identified image file with the results of named entity extraction (including places and characters), the name of the conflict, and their alignment with Wikidata. We also identified sensitive images (marked when the word 'cadavre'/'corpse' appeared in the caption), generated a list of visually and multi-modally similar images (where captions were used to reinforce visual similarity), and detected the presence of humans and/or mounts (horses). For each automatically calculated data item, the 'Eycon' tag was added to track its origin.

(4) RESULTS AND DISCUSSION

(4.1) METADATA ENRICHMENT AND CONTENT EXTRACTION

A multi-modal similarity result was obtained by extracting visual and textual features from the captions (cf. section 3.1). Our approach made it possible to identify subsets (or cliques) of strongly similar photographed events of order 2 to 10 (Figure 4). The subsets were included in the website design and functionality in order to demonstrate the potential application of the dataset for users to explore and compare historical images. Qualitative results are very encouraging, but further research is needed to provide robust metrics on the accuracy of such a method.



clique order	number of cliques
2	19307
3	2902
4	623
5	174
6	69
7	25
8	8
9	6
10	3

Figure 4 Number of images for each clique size.

Human beings and horses were detected across the entire corpus of images and added to the metadata: 71,114 human figures and 8083 horse mounts were recognised out of a total of 77,799 photographs. Many of the objects represented in these decades-old images do not exist anymore and were not labelled in these datasets.

As far as image extraction from albums and newspapers is concerned, metrics were very satisfactory. Approximately 80% of the images were correctly extracted for further use in the project. The other 20% were not detected, but they can be accessed at page level on the online platform. While it is easy to recognise and extract photographs from albums because of their visual features as hand-made objects, trained models are less effective and accurate when it comes to newspapers and magazines. The orderly layouts of some turn-of-the-century English magazines such as the *Illustrated London News* pose few layout segmentation issues, but when it comes to extracting photographs from the layouts of French periodicals such as the *Excelsior*, the pre-trained models come up against photographs inserted in extravagantly shaped frames. A transfer approach with a fine-tuning on public datasets can be considered in future research to improve extraction and classification of detected page regions.

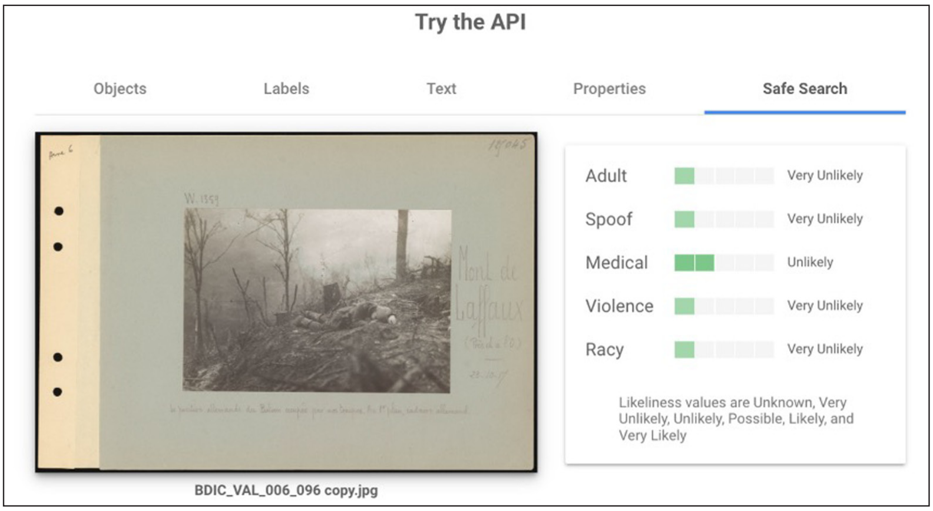
The project team faced difficulties in the extraction of the photographs' captions when they were not already made available in existing metadata. These captions are critical information for historians and archivists. The intricacies of page layouts, the positioning of newspaper captions, and the diverse formats of handwritten album inscriptions significantly increase the difficulty of model training. This complexity remains a major challenge in the analysis of historical documents. A ground truth dataset based on the Forbin/Rumpf collection (Service Historique de la Défense) was nevertheless created with the help of two interns working on the digitization process. They deciphered handwritten captions and stamps for more than a thousand photographs from this early news agency. This dataset will be published and used in future research.

(4.2) ONLINE PLATFORM

The open-access database developed by EyCon aims to rectify and reorient our predominantly Western-centric and Eurocentric perspective on wartime global events. The decision to withhold materials presents an issue: accessibility becomes a privilege when it entails a specialized consultation process inaccessible to all. Dissemination is crucial; digitizing photographic materials and establishing a database could address the concealment of colonial visual records. Both semi-automatic and pre-existing metadata substantiates the EyCon online database that was published via the Omeka S open-source platform for the creation of an online digital library. It is based upon the IIIF format for interoperability of images and metadata. The IIIF framework standardizes image viewing: the user can interact with the image directly on the broadcast platform, but more importantly, metadata is directly attached to the visual medium. In other words, the textual context of the image cannot be easily removed when it is reused. With the same objective, the layout of the online database during document consultation places significant emphasis on text: the metadata contextualizing the photographs are not detached from their visual contents. In addition to displaying the project's results, the website answers questions that might arise from the multiple layers of data collected and created in the course of the project. Photographs with sensitive content are protected by trigger warnings that prevent automatic viewing. However, Omeka S did not meet all our needs despite its flexibility. It is difficult to manage a large visual database with high-resolution images on this publishing platform. Images often take too long to load for a seamless user experience, but efforts have been made to speed up the display time.

(4.3) DETECTING DISTURBING CONTENT

Due to the sensitive nature of the collected corpora, identifying and publishing graphic images turned out to be a challenge (Dentler et al., 2024). Google Vision API tools were tested to identify potentially problematic contents, but the results were inconclusive, as illustrated in Figure 5. Off-the-shelf tools failed to recognise the violent nature of the photograph. 'Safe Search' was designed to automatically detect these materials, but it is not suitable for historical material because the graininess of the image and the lack of context prevent it from detecting potentially disturbing pictures. In some photographs, violence is not visible. It remains implicit: human vision and domain expertise are the only tools one has to read the photograph in its complexity. Context becomes critical to categorize the image as potentially sensitive. An apparently innocuous photograph taken during the Italian invasion of Tripolitania (Libya) in 1911 (Figure 6), which shows what looks like a street parade at first glance, actually depicts a significant historical event. The smiling Italian soldiers in this photograph are seen escorting Arab prisoners to the gallows, where they were about to be publicly hanged for their supposed participation in a rebellion against the colonial power. This is a document of public humiliation and a manifestation of power that can hardly be detected without expert knowledge. It is one of the many examples that illustrates the limitations of elaborate computer vision methods when applied to violent photographic records of painful pasts. In a field where historians and archivists still have difficulties approaching these images in a consensual way, automatic AI tools are rarely suitable in numerous instances.



Adult Very Unlikely
Spoof Very Unlikely
Medical Unlikely
Violence Very Unlikely
Racy Very Unlikely

Likeliness values are Unknown, Very Unlikely, Unlikely, Possible, Likely, and Very Likely

BDIC_VAL_006_096 copy.jpg

Figure 5 Google Vision API test with a photo from the SPA (Valois collection) showing the corpse of a German soldier. 23.10.1917. VAL 006/096. Valois albums – Mont de Laffaux. La Contemporaine.

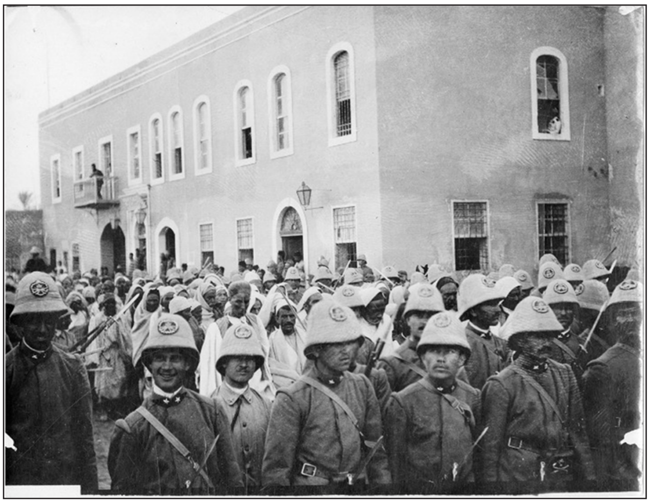


Figure 6 Tripoli – Italian soldiers leading Arabs to their execution, photographic print, 21,5 × 16,5 cm, Victor Forbin collection, courtesy of Service Historique de la Défense (Vincennes, France), access: 2K47160.

The definition of what constitutes a graphic or disturbing image changes over time, making it difficult to create universal and unchanging rules for classifying documents with sensitive content. Several images in the Valois collection – a key subset of EyCon's corpus – illustrate this. This large set of photographs was produced by the photographic section of the French Army during the First World War, which provided the press with images from the front. Contrary to received wisdom, the official visual coverage of the war did not always shy away from the most brutal dimensions of the conflict. Many graphic pictures were thus published in newspapers. At the time of publication, the average reader could access them easily. These disturbing images should now be accompanied by trigger warnings due to their disturbing nature. It was, however, impossible to automatically assign these warnings due to the aforementioned difficulties in detecting violence. Sensitive content was dealt with on a case-by-case basis. While this methodology works well for a limited number of digitized archival images, it cannot easily be applied to massive datasets. More work is needed to establish well-structured ground truth data with a view to augmenting both the historian and the archivist's capacities to process such historically complex data. Semi-automated metadata enrichment remains the best way forward when dealing with violent contents in the photographic archive.

(5) IMPLICATIONS/APPLICATIONS

The showcase website and the published dataset offer various avenues for reusing the data generated by the project. This online digital library allows the public to browse, discover, and use the collections through the available IIIF manifests. Researchers can utilize the ready-to-use dataset to train new models with historical visual content. Given the scarcity of datasets featuring 19th- and early 20th-century photographs suitable for computer vision applications, this dataset serves as a valuable resource. Merging it with other datasets can help expand the study of a longer (and less Western-centric) history of war photography across national contexts, print cultures and photographic memory cultures. The comprehensive metadata, linked to Wikidata, facilitates statistical analysis and geographical mapping of the locations depicted in these images and can be used as a metadata base. Additional multimodal experiments can be conducted on the dataset to correlate visual and textual content. Additionally, the enriched metadata of the photograph corpora, redistributed in the new metadata format to the institutions, will improve content dissemination through enhanced internal search engines accessible to the public. Finally, this project's methodology and workflow can inspire heritage institutions to enhance their metadata collections by analysing visual content.

FUNDING INFORMATION

This research was funded by Idex Université de Paris Cité and a grant jointly funded by the AHRC (Arts and Humanities Research Council, project reference: AH/W008408/1) in the United Kingdom, and the labex "Les passés dans le présent" (Investissement d'avenir, ref. ANR-11-LABX-0026-01) in France.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Marina Giardinetti: Writing – original draft, Data curation

Daniel Foliard: Writing – review & editing, Validation

Julien Schuh: Supervision

Mohamed-Salim Aissi: Software

AUTHOR AFFILIATIONS

Marina Giardinetti orcid.org/0000-0001-9930-5928
LARCA, Université Paris-Cité, Paris, France

Daniel Foliard orcid.org/0000-0001-6400-1801
LARCA, Université Paris-Cité, Paris, France

Julien Schuh orcid.org/0000-0002-0560-5936
MSH Mondes, Université Paris Nanterre, Nanterre, France

Mohamed-Salim Aissi orcid.org/0009-0003-5922-0032
LIP6, Sorbonne Université, Paris, France

REFERENCES

Abgaz, Y., Rocha Souza, R., Methuku, J., Koch, G., & **Dorn, A.** (2021). A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies. *Journal of Imaging, 7*(8), 121. DOI: https://doi.org/10.3390/jimaging7080121

Aissi, M. S. (2023). Comment retrouver des photographies historiques similaires à une image requête? (Master's project report, Sorbonne Université, Master Informatique).

Arnold, T., & **Tilton, L.** (2023). Distant viewing. MIT Press. DOI: https://doi.org/10.7551/mitpress/14046.001.0001

Aske, K., & **Giardinetti, M.** (2023). (Mis)matching metadata: Improving accessibility in digital visual archives through the EyCon Project. *Journal on Computing and Cultural Heritage, 16*(4), 1–20. DOI: https://doi.org/10.1145/3594726

Dentler, J., Jaillant, L., Foliard, D., & **Schuh, J.** (2024). *Sensitivity and access: Unlocking the colonial visual archive with machine learning.* Loughborough University. Retrieved from https://hdl.handle.net/2134/25549789.v1

Ehrmann, M., Hamdi, A., Linhares Pontes, E., Romanello, M., & **Doucet, A.** (2023, September). Named entity recognition and classification on historical documents: A survey. *ACM Computing Surveys, 56*(214), Article 27, 1–47. DOI: https://doi.org/10.1145/3604931

Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., & **Clematide, S.** (2022, August 10). Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents. *Conference and Labs of the Evaluation Forum (CLEF 2022)*. Bologna, Italy. DOI: https://doi.org/10.1007/978-3-031-13643-6_26

Elo, K. (2020). Big data, bad metadata: A methodological note on the importance of good metadata in the age of digital history. In M. Fridlund, M. Oiva & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 103–111). Helsinki University Press. DOI: https://doi.org/10.33134/HUP-5-6

Feng, S., Park, C. Y., Liu, Y., & **Tsvetkov, Y.** (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada. DOI: https://doi.org/10.18653/v1/2023.acl-long.656

Foliard, D. (2020). *Combattre, punir, photographier Empires coloniaux, 1890–1914*. La Découverte. DOI: https://doi.org/10.3917/dec.folia.2020.01

Foliard, D., Schuh, J., Giardinetti, M., Aissi, M. S., & **Dentler, J.** (2024). EyCon project photographs and metadata [Dataset]. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.11449122

Gutehrlé, N., & **Atanassova, I.** (2021). Logical layout analysis applied to historical newspapers. *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*. Silchar, India. https://hal.archives-ouvertes.fr/hal-03468972 (accessed September 23, 2022). DOI: https://doi.org/10.46298/jdmdh.9093

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., & **Weld, D. S.** (2020). The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in Chronicling America. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://dl.acm.org/doi/10.1145/3340531.3412767. DOI: https://doi.org/10.1145/3340531.3412767

Library of Congress. (n.d.). *China and the Boxers: A short history of the Boxer outbreak, with two chapters on the sufferings of missionaries and a closing one on the outlook*. Retrieved from https://www.loc.gov/item/01030948/ (last accessed: 7 June 2024).

Männistö, A., Seker, M., Iosifidis, A., & **Raitoharju, J.** (2022). Automatic image content extraction: Operationalizing machine learning in humanistic photographic studies of large visual archives. *arXiv*. Retrieved from https://arxiv.org/abs/2204.02830 (last accessed: 8 September 2022).

Manovich, L. (2020). *Cultural analytics*. MIT Press. DOI: https://doi.org/10.7551/mitpress/11214.001.0001

Moretti, F. (2013). "Operationalizing": Or, the function of measurement in modern literary theory. *New Left Review, 84*. https://newleftreview.org/issues/ii84/articles/franco-moretti-operationalizing

Rameau. (n.d.). *Boxers, révolte des (1899–1901)*. Retrieved from https://data.bnf.fr/fr/12070712/chine_--_1899-1901__revolte_des_boxeurs_/#linked_rameau_broader (last accessed: 7 June 2024).

Schill, P. (2024). The brutalised bodies of a colonial conquest before the court of global opinion: Photography, media uses, and emotions during the Italo-Turkish war in Tripolitania (1911–1912). *History of Photography*. Routledge. (Accepted for publication).

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & **Li, W.** (2021). LayoutParser: A unified toolkit for deep learning based document image analysis. *arXiv*. Retrieved from https://arxiv.org/abs/2103.15348 (last accessed: 20 June 2024). DOI: https://doi.org/10.1007/978-3-030-86549-8_9

Wevers, M., & **Smits, T.** (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities, 35*(1), 194–207. DOI: https://doi.org/10.1093/llc/fqy085

Zhang, Z., Li, J., Stork, D. G., Mansfield, E., Russell, J., Adams, C., & **Wang, J. Z.** (2022). Reducing bias in AI-based analysis of visual artworks. *IEEE BITS The Information Theory Magazine, 2*(1), 36–48. DOI: https://doi.org/10.1109/MBITS.2022.3197102

TO CITE THIS ARTICLE:
Giardinetti, M., Foliard, D., Schuh, J., & Aissi, M.-S. (2024). The EyCon Dataset: A Visual Corpus of Early Conflict Photography. *Journal of Open Humanities Data*, 10: 40, pp. 1–12. DOI: https://doi.org/10.5334/johd.213

Submitted: 03 April 2024
Accepted: 07 June 2024
Published: 04 July 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.