Image Retrieval in Digital Libraries

A Large Scale Multicollection Experimentation of Machine Learning techniques

Jean-Philippe Moreux
Preservation dpt, Digitization service, Bibliothèque nationale de France, Paris, France.
jean-philippe.moreux@bnf.fr
Guillaume Chiron
L3i Lab, University of La Rochelle, France
guillaume.chiron@univ-lr.fr

Abstract: While historically digital heritage libraries were first powered in image mode, they quickly took advantage of OCR technology to index printed collections and consequently improve the scope and performance of the information retrieval services offered to users. But the access to iconographic resources has not progressed in the same way, and the latter remain in the shadows: manual incomplete and heterogeneous indexation, data silos by iconographic genre. Today, however, it would be possible to make better use of these resources, especially by exploiting the enormous volumes of OCR produced during the last two decades, and thus valorize these engravings, drawings, photographs, maps, etc. for their own value but also as an attractive entry point into the collections, supporting discovery and serenpidity from document to document and collection to collection. This article presents an ETL (extract-transform-load) approach to this need, that aims to: Identify and extract iconography wherever it may be found, in image collections but also in printed materials (dailies, magazines, monographies); Transform, harmonize and enrich the image descriptive metadata (in particular with machine learning classification tools); Load it all into a web app dedicated to image retrieval. The approach is pragmatically dual, since it involves leveraging existing digital resources and (virtually) on-the-shelf technologies.

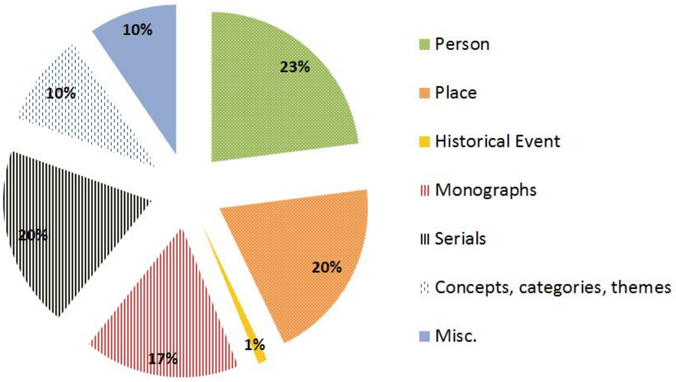
Keywords: digital libraries; image retrieval; CBIR (content based image retrieval) ; automatic image classification; machine learning; data mining; metadata; OCR; heritage documents

1 Introduction

Even though the creation of digital heritage collections began with the acquisition in image mode, several decades later to search in the content of some of these images still belongs to a more or less distant future [Gordea16]. This apparent paradox originates in two facts: (1) the massive OCR processing of printed materials has rendered major services in terms of information retrieval; (2) searching or browsing large collections of images remains a challenge, despite the efforts of both the scientific community and GAFAs to address the underlying challenges [Datta08].

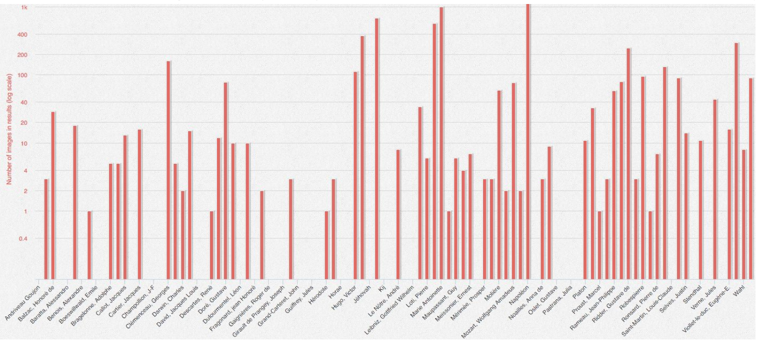
In digital libraries, however, the needs are very real, if one believes user surveys (63% of the users of Gallica consult images, 85% know the existence of an images collection [BnF 17]) or statistical studies of user behavior: of the Top 500 most common queries, 44% contain entities named (Person, Place, Historical Event, see Fig. 1 [Chiron17]), for which we can advance that iconographic resources would provide information complementary to the ones present in the textual content.

Figure 1: Distribution of the Top 500 user queries by type (gallica.bnf.fr, Dec 2015-March 2016, 28M queries)



In order to answer these encyclopedic-liked queries (see Fig. 1: 90% of the Top 500 queries target either named entities, title works or concepts), digital libraries are not without resources. Fig. 2 shows the number of images provided by Gallica's iconographic collection (photographs, engravings, posters, maps, etc.) for the Top 100 queries on Person NE.

Figure 2: Images retrieved (using dc:subject) for the Top 100 queries on a named entity of type Person



It also highlights gaps related to its relatively small size (≃1.2M images), given the broad spectrum of areas of knowledge and periods surveyed by users (from Antiquity to the 21th century). However, libraries are rich in many other iconographic sources (e.g. the newspapers, with up to three illustrations per page on average for the most illustrated titles of the first half of the XXth c. [Moreux16]). But sometimes organized in silos of data that are not interoperable, most often lacking the descriptors indispensable to image search, whether they are "low level" (size, color, shape) or semantics (thematic indexing, annotation[1], etc.). And when all the documentary genres are gathered within a single portal[2], the search and browsing methods will have been generally designed according to a classical paradigm (bibliographic metadata and full-text indexing; page flip). While the querying of iconographic content poses special challenges (see [Picard15]), corresponds to various uses (from playful mining of old photographs to serious study of illuminated or printed materials of manuscripts or incunabula [Coustaty11]), and targets different knowledge domains (encyclopedic or specialized), and calls for specific human-machine interactions (from basic keyword input to the drawing of a sketch representing the image being searched; see [Gang08], [Breiteneder00], [Datta08]).

This article presents a proposal for a pragmatic solution to these two challenges, the creation of an encyclopedic image database (covering several collections of a digital library), and of its interrogation methods. A first section describes the initial phase of the necessary aggregation of the heterogeneous data and metadata available (the *e* of an ETL –Extract-Transform-Load– approach, shown schematically in Fig. 3). The second section presents the transformations and enrichments applied to the collected data, in particular the application of treatments under the so-called "machine learning" methods. Finally, a multimodal query mode is tested and its results are commented on.

Figure 3: ETL process and its tools

1. Extract	**2. Transform**	**3. Load**
Gallica APIs	Watson Visual Recognition API (IBM)	BaseX
OAI-PMH	TensorFlow Inception-v3 model (Google)	XQuery
SRU	IIIF API	IIIF API
	Tesseract	Mansory.js
Perl, Python		

2 Extract and Aggregate

Several decades after the creation of the first digital heritage libraries (Gallica celebrates its 20th anniversary in 2017), the iconographic resources preserved in the digital stores are both

[1] Manual annotation is not a universal remedy (see, for example, [Nottamkandath14], [Welinder10]).

[2] This is the case of Gallica (http://gallica.bnf.fr).

massive and constantly expanding. Mediation[3,4] and manual indexing actions[5] have been carried out but their cost limits them to restricted and generally thematic and/or homogeneous collections (photography, poster, illumination, etc.). On the other hand, a massive multicollection approach requires a first step of aggregation of content in order to take into account the variability of the data available, due both to the nature of the documentary silos and to the history of the digitization policies that have influenced their constitution.

The image database described in this article aggregates 340k illustrations (for 465k pages) of the Gallica's collections of images and prints relating to the First World War (1910-1920 time period). It follows an XML formalism and has been loaded into an XML database[6] using the Gallica APIs[7], SRU and OIA-PMH protocols. Its data model (see Fig. 4, Appendix) aggregates document, page and illustration levels; it allows to receive the information available in the different documentary silos targeted (the distribution of which is given in the appendix, Fig. 5). Access to the illustrations themselves is carried out with the IIIF Image 2.0 API.

The feedback from this multicollection aggregation step is presented in the following sections. But at a first glance, we can say that this first step well worth the pain because it gives access to invisible illustrations to users. Nevertheless, some challenges exist: heterogeneity of formats and metadata available; computationally intensive (but parallelizable); noisy results for newspapers.

2.1 Images Collection

A pre-existing thematic set of the OAI-PHM warehouse of the digital library is used to extract the metadata of 6,600 image documents (graphic works, press clippings, medals, cards, musical scores, etc.) resulting in a collection of approximately 9,000 illustrations. These documents present particular challenges: metadata suffering from defects of incompleteness and inconsistency (due to the variability of indexing practices); little or no image metadata (genre: photo, engraving, drawing… ; color and size of the original document); portofio (e.g. Fig. 9: cover and blank/text pages must be excluded). This corpus (see appendix, Fig. 6) was supplemented by various SRU requests on catalog metadata ("Subject = War 14-18", "source = Meurisse photo agency", "type = poster").

2.2 Printed Collection

The database is fed by an bibliographic selection of books and magazines as well as by a temporal sampling of the newspapers collection. Here, the OCRed text surrounding the illustration is extracted and preserved as a textual descriptor.

[3] Europeana: http://blog.europeana.eu/2017/04/galleries-a-new-way-to-explore-europeana-collections

[4] BnF: http://gallica.bnf.fr/html/und/images/images

[5] British Library: https://imagesonline.bl.uk

[6] BaseX: http://basex.org

[7] https://github.com/hackathonBnF/hackathon2016/wiki

2.2.1 Newspaper and magazines

The BnF digital serials collection is presented under different formalisms related to the history of successive digitization projects. In all cases, it is a question of extracting the descriptive metadata from the METS and ALTO[8] formats. In the case of the recent digitization projects identifying the articles structure (OLR, optical layout recognition), this task is facilitated because of their fine and controlled structuring; the oldest programs offer raw OCR with little structure. Thematic journals enrich the database: trench newspapers, scientific and technical journals, military science journals, etc.

In the case of the daily press, the illustrations are characterized by singularities (variable size of illustrations, from double spread page to thumbnail portraits; poor reproduction quality, especially at the beginning of photogravure); a wide variety of types (from map to comic strip) and a large volume (Fig. 7, Appendix).

Noise is also massive (blocks of text mistakenly recognized by the OCR as illustrations; ornaments; illustrated advertisements repeated throughout the publications). Various heuristics are applied to reduce this noise: filtering on physical criteria (size; ratio width to height to remove dividing lines); location of the illustrations (headings of the first page; last page containing advertisements). This step leads to 271k usable illustrations (on 826k collected, ie a noise of 67%). A second filter to identify the residual spurious text ads and text blocks is carried out in a later step (see section 3.3.1). Note that a image search by similarity could also filter the recurring advertisements and illustrated headers.

2.2.2 Monographs

The same treatment is applied to the OCR of the monographs corpus (Fig. 10, appendix): historical books, history of regiments, etc.

3 Transform and Enrich

This step consists of transforming, enriching and aligning the metadata obtained during the aggregation phase. Indeed, the descriptive metadata of the collected illustrations are characterized both by their heterogeneity (in extreme cases, several hundred illustrations are placed under a single bibliographic record) and by their poverty regarding the expected user functionalities.

3.1 Text Extraction

Illustrations of printed materials without text descriptor (due to a missed text block in the original OCR, see Fig. 22, on the right) are detected and their enlarged bounding box is processed by the Tesseract OCR engine, allowing textual indexing of those “silent” artwork.

[8] The British Library Mechanical Curator is one of the sources of inspiration for this exotic use of the OCR (http://mechanicalcurator.tumblr.com).

3.2 Topic Extraction

3.2.1 Images Collection

The IPTC[9] thematic indexing of press contents is carried out using a semantic network approach: the keywords of the documents record (title, subject, description, format...) and illustration captions (if any) are lemmatized and then aligned with the IPTC topics. Such a method is not easily generalizable (the network has to be refined manually according to the corpus). On a reduced corpus, however, it allows to offer a rudimentary but operative classification.

3.2.2 Printed materials Collection

On the contrary, the printed materials are characterized by a rich textual apparatus (title and legend, text preceding or following the illustration) that is possible to topicalize. Various techniques for the detection of topics would be operational here[10] (see, for example, the topic modeling method LDA without supervised learning [Underwood12], [Langlais17], [Velcin17]). On news, which is a polyphonic media in essence, this topic modeling is unavoidable. Press corpora that are digitized with article separation (OLR) sometimes include partial topic characterization, usually done manually by digitization providers (e.g. classified ads, advertisements, stock exchange, judicial chronicles), which can be included into the metadata related to topic classification. Let us note that content from some thematic magazines (sciences, sports...) are also assignable to an IPTC theme.

3.3 Extracting Metadata from Images

The search in image contents faces a twofold gap: on one hand, between the reality of the world recorded in a scene (in our context an “illustration”) and the informational description of this scene; on the other hand, between the interpretations of a scene by different users (possibly following different research objectives). Reducing or overcoming these gaps (sensory and semantic) implies to provide operational descriptors (nature of illustrations, color, size, texture, etc.) to the system as well as to their users. This enables the search to be operated in a space shaped by these visual descriptors. Quality is also a criterion to be taken into account, which by nature is however difficult to quantify. Taking the case of heritage photographs, a distinction should be made between silver print photography and any other reproduction methods.

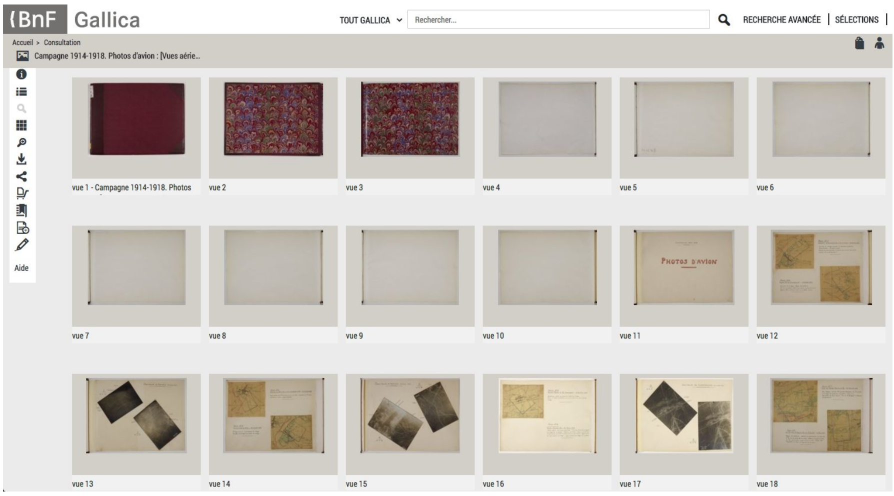
3.3.1 Image Genres Classification

The genre of illustrations (e.g. photography, line drawing, engraving...) is not always characterized in the catalog records (Fig. 9). Of course, this information is also not more available over illustrations of printed materials.

[9] http://cv.iptc.org/newscodes/mediatopic, 17 top-level topics.

[10] Task not performed during this experiment.

Figure 9: Unknown genre-illustrations in portfolio: map, photo, sketch...



To overcome this lack, a deep-learning based method for image genres classification is implemented.

Modern neural-networks-based approaches (see [Pourashraf15] for a SVM approach of images genre classification) are able to recognize a thousand different common objects (e.g. boat, table, dog...) and surpasses human-level performance on some datasets. For exemple, the Inception-v3 [Christian15] is the third iteration of improvement over the original GoogLeNet model (a 22 layers convolutional neural network that won the ILSVRC 2014 challenge). This kind of models are usually pre-trained on supercomputers and are specially optimized to perform well on common dataset such as ImageNet [FeiFei10] (a major reference in the field in terms of size and representativeness).

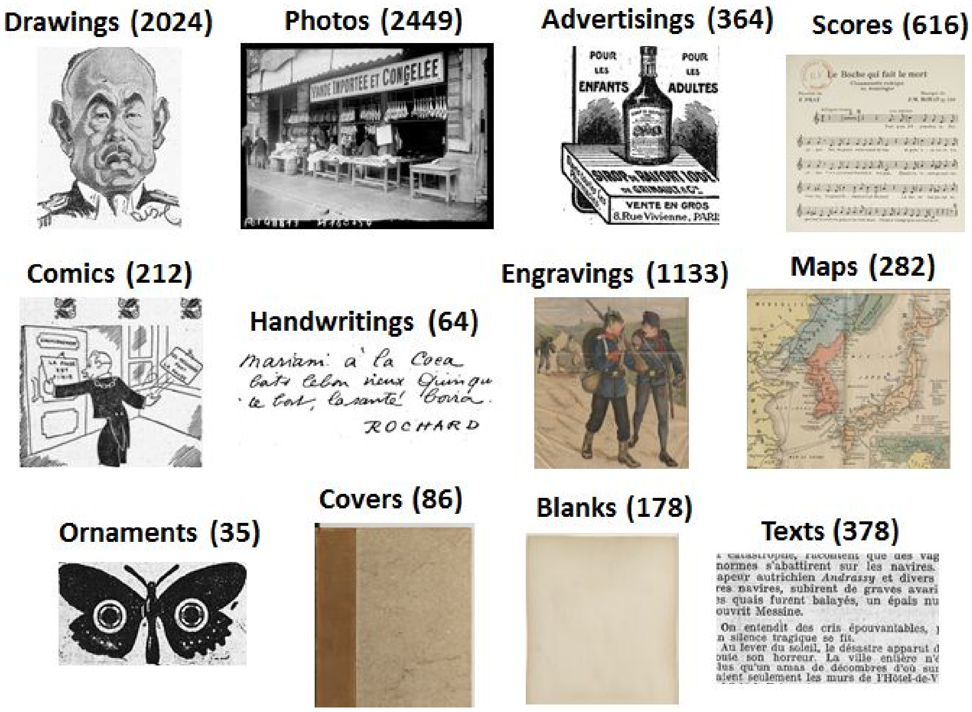
The ever-increasing effort to improve these models benefits the "computer vision" community in general but not only. Indeed, it is now possible to take advantage of the power capitalized by these models over other problems, the classification of heritage documents in our case. This can be done by retraining a subpart of the model, basically the last layer (which takes a few hours on a conventional computer, compared to the months that would be required to retrain the full model) following the so-called "transfer learning" approach [Pan09]. The "transfer learning" consists in re-using the elementary visual features found during the original training phase (as they have shown their potential to classify a given dataset), but on a new custom dataset with the expectation that those given features would still perform a good classification on that new dataset. Also, reducing the number of classes (which somehow simplifies the targeted problem) helps to keep honorable classification scores, even though the model was originally not trained specifically for the task in question.

Figure 10 gives an overview of the twelve genre categories our Inception-v3 model have been trained on (drawing, photo, advertising, musical score, comic, handwriting, engraving, map, ornament, cover, blank page and text, extracted from all the collections). The training has required documents labeled by their class (7,786 in our example). Once trained, the model is then evaluated using a test dataset (1,952 documents). Figure 11 details the results obtained.

The global accuracy[11] is 0.90 (computed over all the classes) with a recall[12] of 0.90 which corresponds to a similar F-measure[13] of 0.90. These results are considered to be good regarding the size and the diversity of the training dataset, and performances can be better with less generic models (on monograph and image collections only, F-measure is 0.94) or a full trained model (but at the cost of the calculation time).

This model is also used to filter unwanted illustration genres: noisy ornament and text blocks from newspapers (and eventually the illustrated ads), covers and blank pages from portofolios (see Section 2.1). A full-scale test (6,000 illustrations) on a newspaper title[14] without usable images (but ads) leads to a 0.98 global recall rate for filtering noisy illustrations.

Figure 10: The twelve categories of the training dataset (*number of documents are given for each class)*



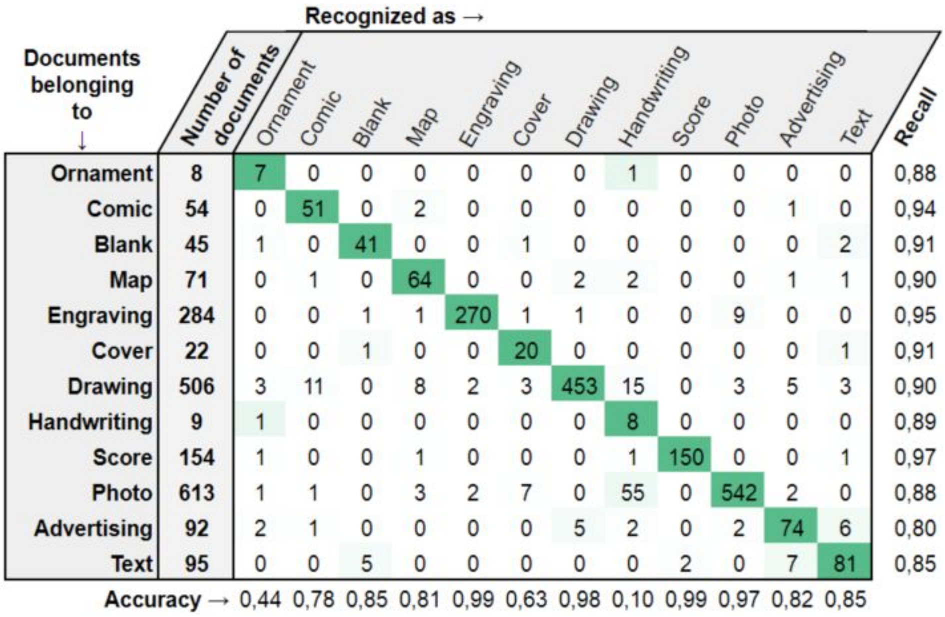
[11] Number of relevant documents found in relation to the total number of documents found by the classifier.

[12] Number of relevant documents found by the classifier in terms of the number of relevant documents held by the database.

[13] Measure that combines precision and recall.

[14] *Le Constitutionnel*, http://gallica.bnf.fr/ark:/12148/cb32747578p/date

Figure 11: Classification results over the twelve genre categories



		Recognized as →												
Documents belonging to ↓	Number of documents	Ornament	Comic	Blank	Map	Engraving	Cover	Drawing	Handwriting	Score	Photo	Advertising	Text	Recall
Ornament	8	7	0	0	0	0	0	0	1	0	0	0	0	0,88
Comic	54	0	51	0	2	0	0	0	0	0	0	1	0	0,94
Blank	45	1	0	41	0	0	1	0	0	0	0	0	2	0,91
Map	71	0	1	0	64	0	0	2	2	0	0	1	1	0,90
Engraving	284	0	0	1	1	270	1	1	0	0	9	0	0	0,95
Cover	22	0	0	1	0	0	20	0	0	0	0	0	1	0,91
Drawing	506	3	11	0	8	2	3	453	15	0	3	5	3	0,90
Handwriting	9	1	0	0	0	0	0	0	8	0	0	0	0	0,89
Score	154	1	0	0	1	0	0	0	1	150	0	0	1	0,97
Photo	613	1	1	0	3	2	7	0	55	0	542	2	0	0,88
Advertising	92	2	1	0	0	0	0	5	2	0	2	74	6	0,80
Text	95	0	0	5	0	0	0	0	0	2	0	7	81	0,85
Accuracy →		0,44	0,78	0,85	0,81	0,99	0,63	0,98	0,10	0,99	0,97	0,82	0,85	

3.3.2 Size, Color and Position

When the “color mode” information is not provided by the scan metadata, it can be extracted from each illustration. The originally monochromatic documents (black and white, sepia, selenium, etc.) scanned in color are a problematic case where a naive approach based on the hue components of the HSV model can be used (see also Section 3.4.3).

The position, size and density of illustrations per page are also extracted. In the case of the newspapers, searching either for the front page or for a large illustration (see Fig. 12, Appendix) is a common and legitimate need.

3.4 Extracting Content from Images

Historically (see [Datta08]), content based image retrieval (CBIR) systems were designed to: 1) extract visual descriptors from an image, 2) deduce a signature from it and 3) search for similar images by minimizing the distances into the signatures space. The constraint that CBIR systems can only by queried by images (or signatures) has a negative impact on its usability [Gang08]. Moreover, it appeared that similarity measures struggled to encode the semantic richness and the subjectivity of interpretation of image contents, despite the improvements brought to CBIR over the time (e.g. considering subregions of an image).

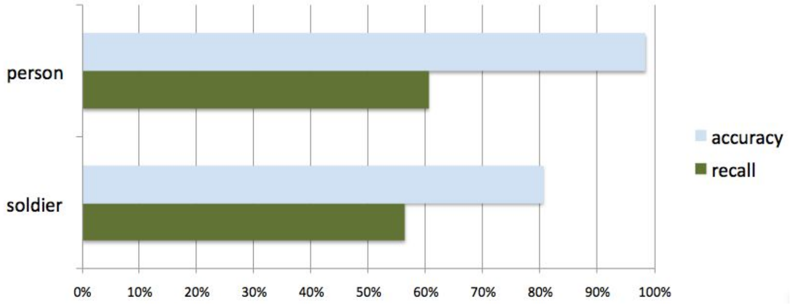
In recent years, advances in deep learning techniques tend to overcome these limitations, in particular tanks to clustering and classification (or concept extraction) approaches, the latter offering the possibility of generating textual descriptions from images, and thus supporting textual queries [Karpathy17]. The IBM Visual Recognition API (being part of Watson IA services[15]) illustrates these evolutions[16]. The following sections describe its application over our collection of WW1 heritage images.

3.4.1 Concepts Detection

The Visual Recognition API relies on deep learning algorithms to analyze the images and extract different concepts (objects, people, colors, etc.) which are identified within the API classes taxonomy. The system returns pairs of estimated "class/confidence".

Here is described an evaluation carried out on person detection. A Ground Truth (GT) of 2,200 images is created, covering several representative image genres of the database (photo, engraving, drawing). The distribution of illustrations with and without person is set to 80/20, which is also representative of the collection. Then, another evaluation is conducted on the "Soldier" class (600 images, with a 50/50 distribution).

Figure 13: Recall and accuracy for the "Person" and "Soldier" classes detection



The class "Person" has a modest recall of 60.5%[17] but benefits from excellent accuracy of 98.4% over the 1190 illustrations provided to the users. A decrease is observed for the more specialized class "Soldier" (56% recall and 80.5% accuracy). However, these results are to be compared with the relative silence of the classical approaches: the concept "Person" does not exist in the bibliographic metadata (a fortiori on non-cataloged newspapers illustrations!). A search on the keyword "person" in the GT returns only 11 correct illustrations. Analogously the keyword "soldier" returns 48 results. Therefore, it would be necessary to write a complex request like "soldier OR military officer OR gunner OR…") to obtain a 21% recall, to be compared to the 56% obtained by using the visual recognition approach. It should be noted that a quite interesting 70% recall is obtained when both query modes are mixed together.

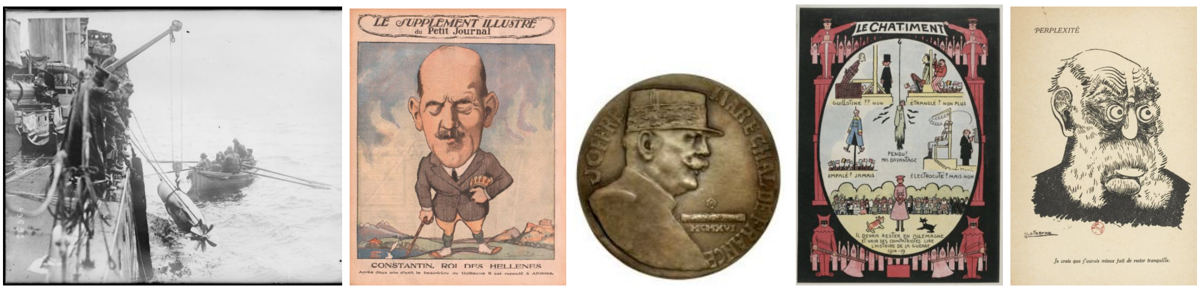
The recalls for the class "Person" over different documentary genres is analysed: engraving and drawing: 54%, silver photo: 67%, photogravure: 72%. It should be noted that the API gives results even on "difficult" documents (Fig. 14).

[15] https://www.ibm.com/cognitive

[16] Google TensorFlow Object Detection API should also have been used.

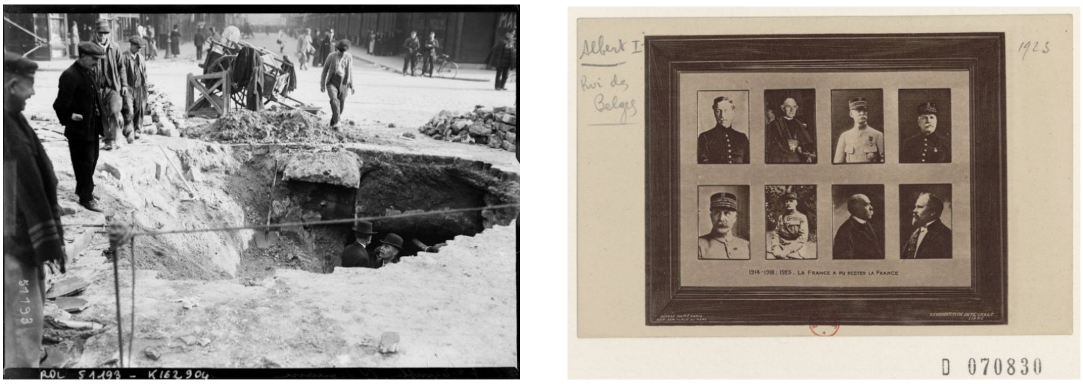
[17] A less strict GT, excluding for example blurry silhouettes or too small thumbnails, would lead to a better recall.

Figure 14: Samples of results for Person recognition


LE SUPPLÉMENT ILLUSTRÉ du Petit Journal
CONSTANTIN, ROI DES HELLENES
JOFFRE MARECHAL DE FRANCE
MCMXVI
LE CHATIMENT
PERPLEXITÉ

The photogravure genre has a higher recall than the silver photo one, which may seem surprising, but can be explained by the complexity of the scenes in the image collection compared to the illustrations of printed materials (simpler scenes, smaller formats). Generally, complex scenes (multi-objects) highlight the current limitations of these technologies and the need to overcome them (see, for example, a generative model of textual descriptions of images and their subregions [Karpathy17]). Figure 15 shows such an example, as well as another tricky case, portraits in a picture frame, which is classified as a picture frame (and not as persons).

Figure 15: Complex scene: the API suggests “explosive device” and “car bomb” but not “person” (left); picture frame (right)



3.4.2 *Face Detection*

The “Gender studies” are a full-fledged field of research. Also, the reuse of digital visuals of human face for recreational [Feaster16] or scientific [Ginosar15] purposes has its followers. It is therefore not insignificant for a digital library to take into account such needs. The Watson API offers a face detection service, which also provides the estimated age and gender (M/F) of the detected persons. Figure 16 shows that detecting faces (M and F) is achieved with a 30% recall and an accuracy close to 100%. The corpus is considered difficult for this kind of task as it includes drawings, engravings, degraded photos, etc. (see Figure 17). An even lower recall of 22% is observed for the Male/Female detection task, and there is especially a poor accuracy of 26.5% for the “Female” class (the API tends to populate the world with "moustached" women...). Imposing a 50% threshold on the confidence estimate (parameter provided by the API), the accuracy for the Female class improves but to the detriment of the recall.

Figure 16: Recall and accuracy over the classes Face, Male and Female

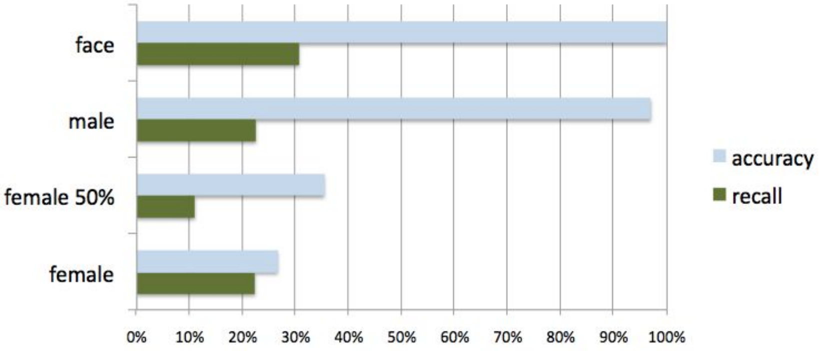
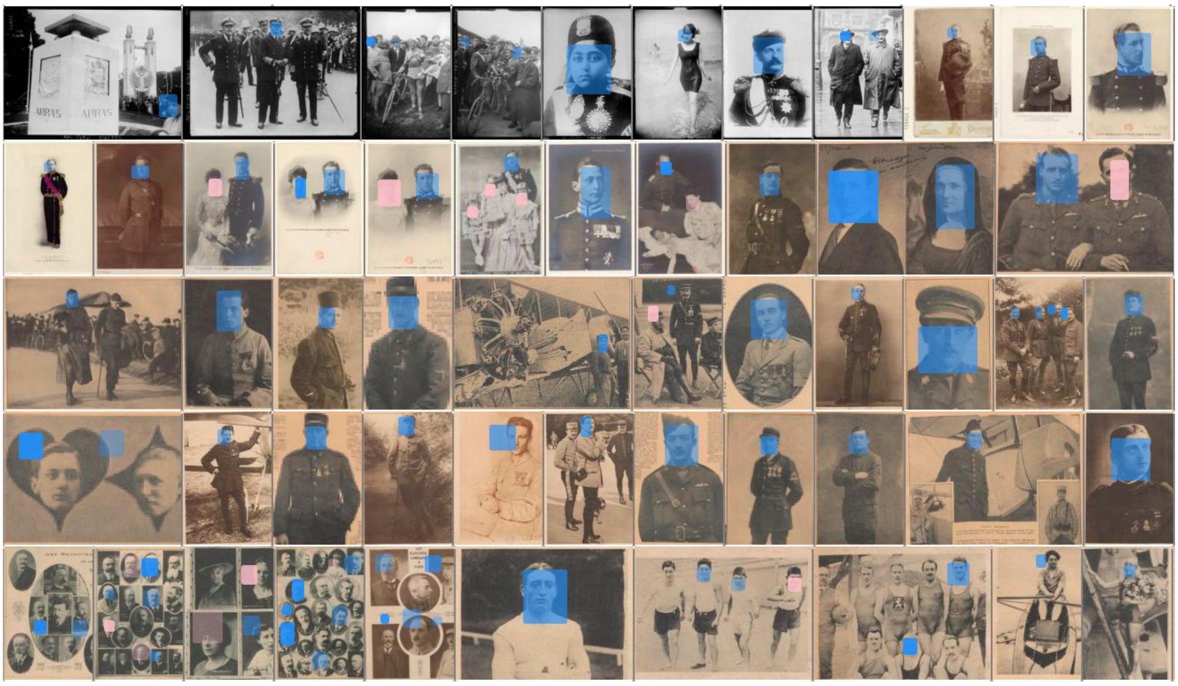


Figure 17: Example of faces detected by the API

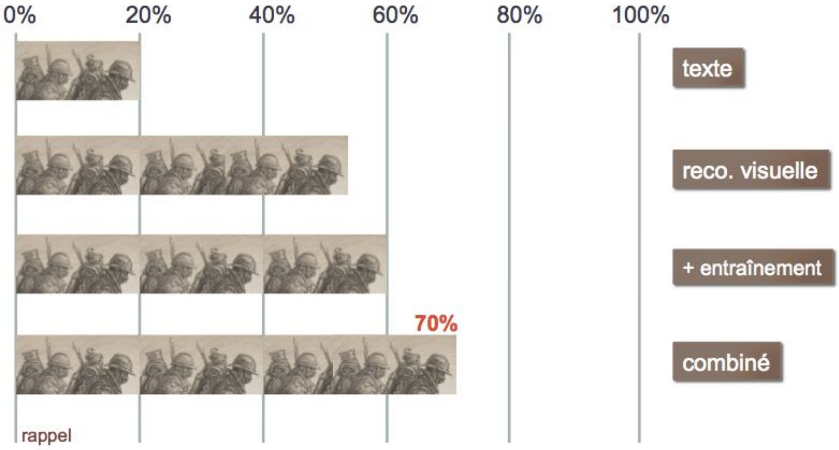


The cumulative use of the outputs from the two recognition APIs (Person and Face Detection) results in an improvement of the overall recall for person detection to 65%.

The API also supports the creation of supervised classifiers. It works by providing a training corpus (e.g. collection of images labeled as Person (Man, Woman) / Non-person), after what a split of the GT is reanalyzed. This experience provided a significant improvement on the person detection task (65% recall and 93% accuracy) but had only a small effect on the gender detection tasks. The overall recall (using both the generic API class and the ad hoc classifier combined) is also improved with a final rate of 85%.

Figure 18 summarizes the recall rates for the Soldier class according to the four interrogation modalities analysed (textual descriptors, visual recognition, visual recognition with classifier, combined text + visual) and shows the obvious interest in offering users a search multimodal.

Figure 18: Soldier class recall rates for the 4 interrogation modalities



3.4.3 *Color Detection*

The color classes provided by the API (one or two dominant colors per image) can be made available to the users in order to query the image base on this criterion (see Fig. 19, Appendix). Finally, let's notice that the Watson API also offers a search over similarities functionality[18] (not evaluated in this work).

4 Loading and Interacting

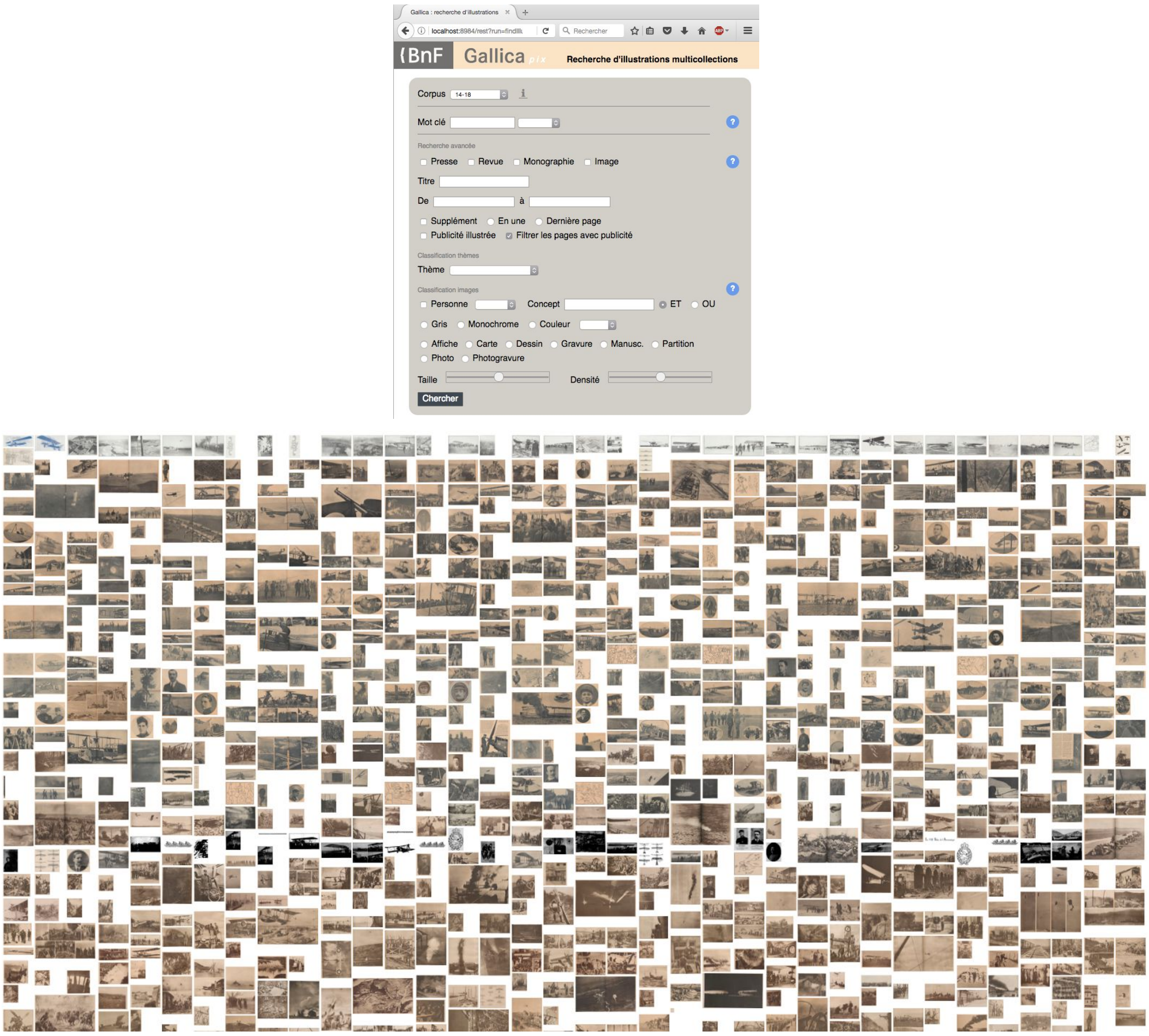
XML metadata is loaded into a BaseX database made accessible through REST queries (client/server mode). Users can create requests using XQuery/FLOWR expressions and submit them with a HTML form. The mosaic of images is created with the JS Mansory library and fed on the fly by the Gallica IIIF server. A rudimentary faceted browsing functionality (color, size, genre, date) allows to prefigure what a successful user/system interaction would be.

The complexity of the form and the large number of results (see Fig. 20) it often leads to reveal, if need be, that searching and browsing in image databases carries specific issues of usability and remains a subject of research topic in its own right (see for example [Lai13]). Thus, the operational modalities of multimodal querying on illustrations (in the sense of [Wang16]: by image content and bibliographic or OCR textual descriptors) must be made intelligible to the users. Also the presence of false positives and noise in the results provided (but this landscape is close to that of the OCR, which is now familiar to users of digital

[18] See https://bildsuche.digitale-sammlungen.de for a real-life large scale implementation of similarity search.

libraries). However, this type of research contributes to narrow the gap between the formulation of the user need (e.g. “a good quality classroom picture in 1914”) and how the data are understood by the system. The following details examples of usage that quite well represent the usual queries that can be submitted to a database of encyclopaedic heritage images.

Figure 20: Image retrieval form; 1,000 images result example (below)



Gallica : recherche d'illustrations

localhost:8984/rest?run=findIll

Rechercher

BnF Gallica pix Recherche d'illustrations multicollections

Corpus 14-18

Mot clé

Recherche avancée

□ Presse □ Revue □ Monographie □ Image

Titre

De à

□ Supplément ○ En une ○ Dernière page

□ Publicité illustrée ☑ Filtrer les pages avec publicité

Classification thèmes

Thème

Classification images

□ Personne Concept ● ET ○ OU

○ Gris ○ Monochrome ○ Couleur

○ Affiche ○ Carte ○ Dessin ○ Gravure ○ Manusc. ○ Partition

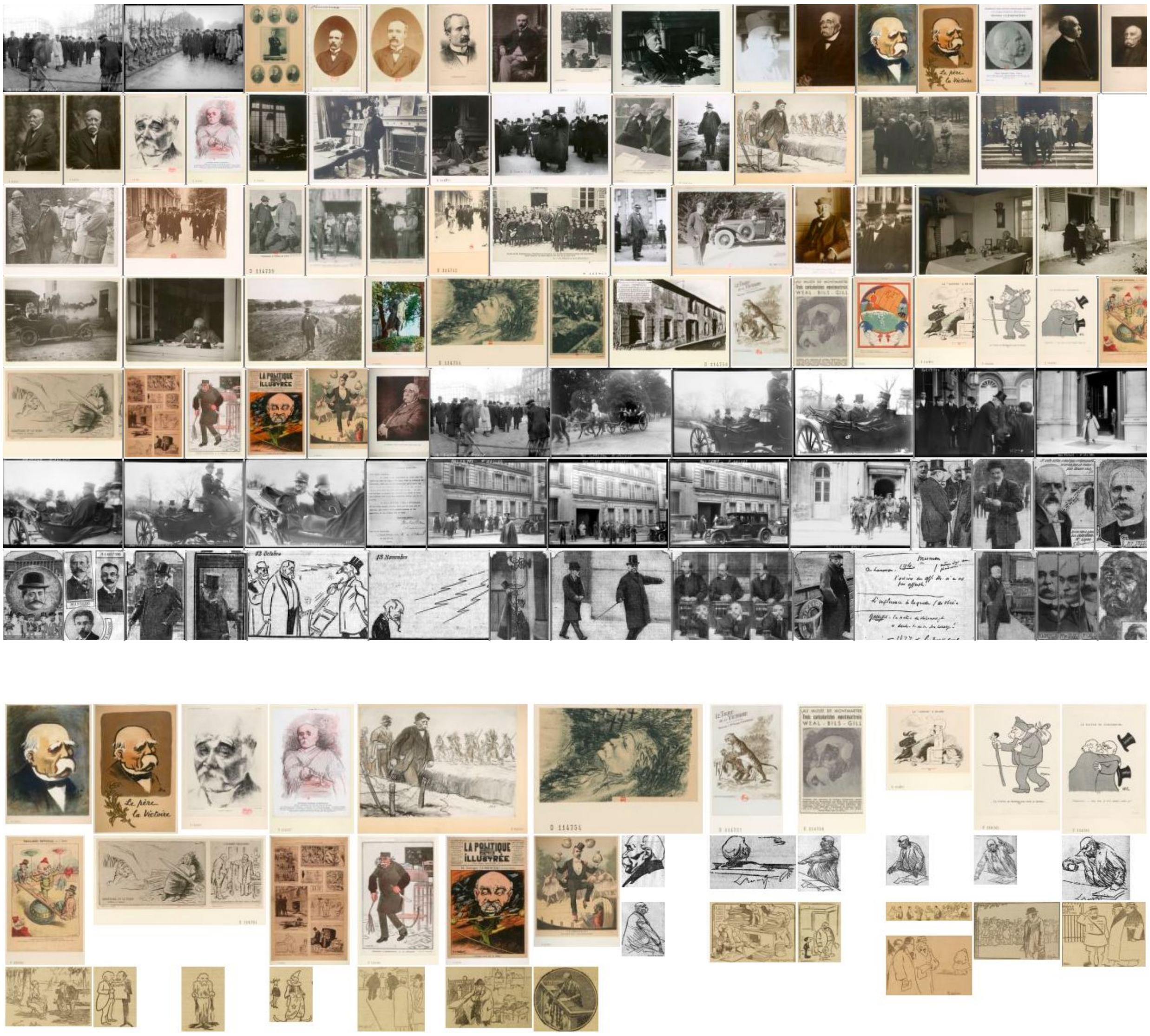
○ Photo ○ Photogravure

Taille Densité

Chercher

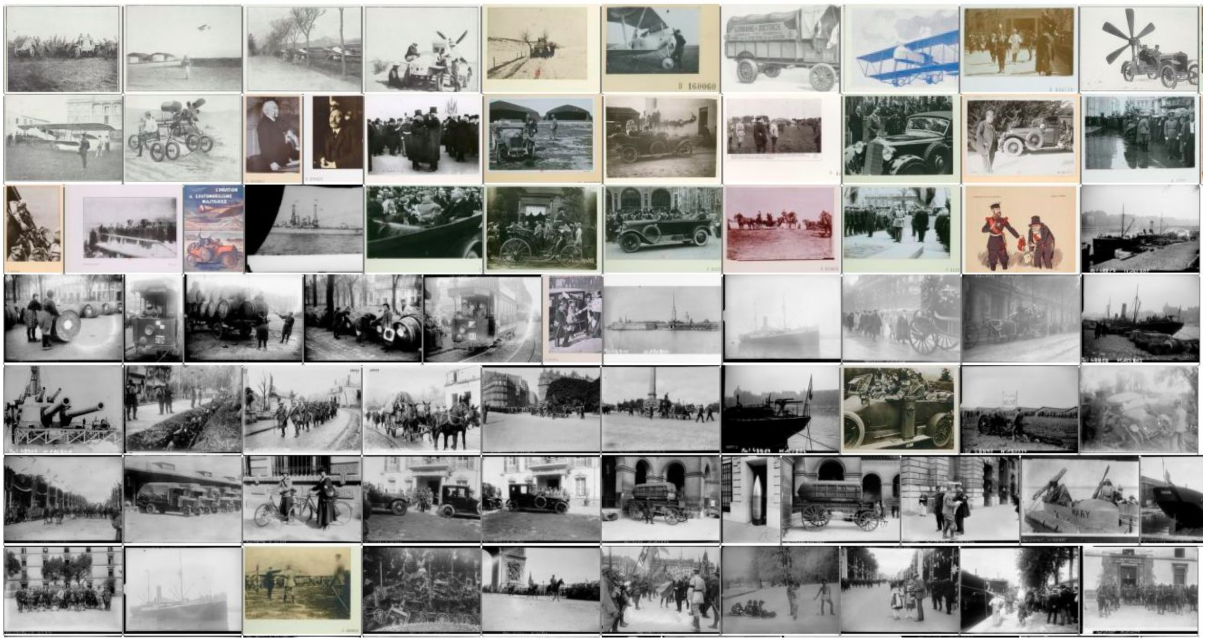
Encyclopedic query on a named entity: textual descriptors (metadata and OCR) are used. Among the Top 100 person-type queries submitted to Gallica (see Section 1), one is related to “Georges Clemenceau” (130 results). The same query now returns more than 1,000 illustrations with a broader spectrum of image genres. The faceted browsing can then be applied by users to refine the search (e.g. Clemenceau caricatures in dailies can be found with the “drawing” facet, Fig. 21). Here, the accuracy/recall rates are correlated to the quality of the textual descriptors.

Figure 21: Mosaic (samples) returned over the query “Clemenceau”; below, caricatures filter



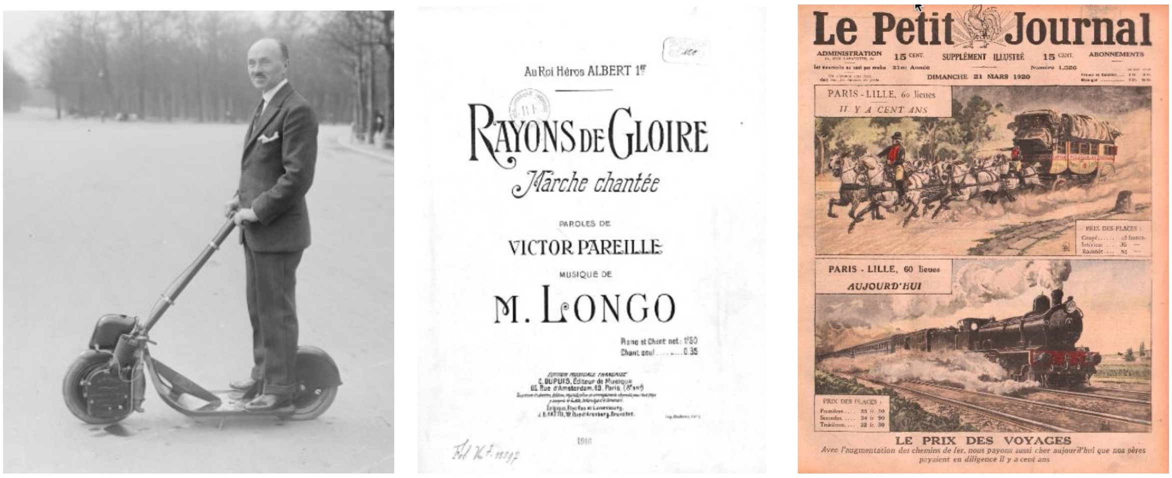
Encyclopedic query on a concept: the conceptual classes extracted by the Watson API overcome the silence either related to the bibliographic metadata or to the OCR but also circumvent the difficulties associated with multilingual corpora (some documents in the database are of German origin) or the lexical evolution (see Fig. 26, Appendix). The IPTC topics (or any other content indexing system) emerge from the same use case. In the context of the Great War, it’s easy to think for example about persons (cf. supra: genders, soldiers...), vehicles, weapons, etc. In this case, the user does generally not expect completeness and accuracy but rather “suggestions”. Figure 22 shows the example of a query on the superclass “vehicle”, which returns many instances of its subclasses (bicycle, plane, airship, etc.).

Figure 22: Query “Class=Vehicule” (sample)



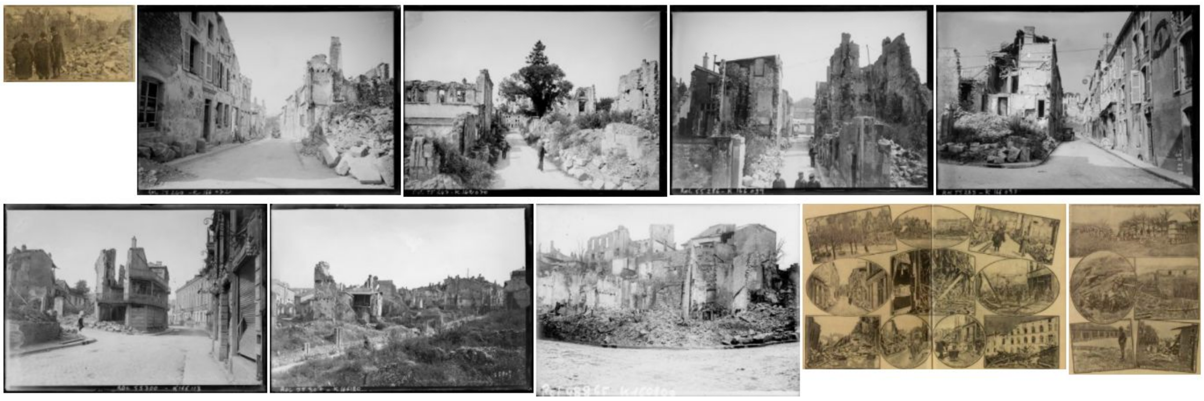
The effects of machine learning are sometimes felt, and in particular those related to the underlying process of generalization. Namely, this 1917 motorized scooter is labeled as a “Segway” (see Fig. 23, left) and this music score title page (middle) is indexed as a bourgogne wine label. On the contrary, the illustration on the right representing a steam locomotive returns to light under the denomination of “armored vehicle”. Let us keep in mind that machine learning techniques remain dependent on the modalities over which the training corpus has been created [Ganascia17]. The most advanced of them which are for example trained to recognize dogs playing frisbee [Karpathy17], will not be necessarily be advantaged on documents of the early twentieth century…

Figure 23: Deep learning artifacts



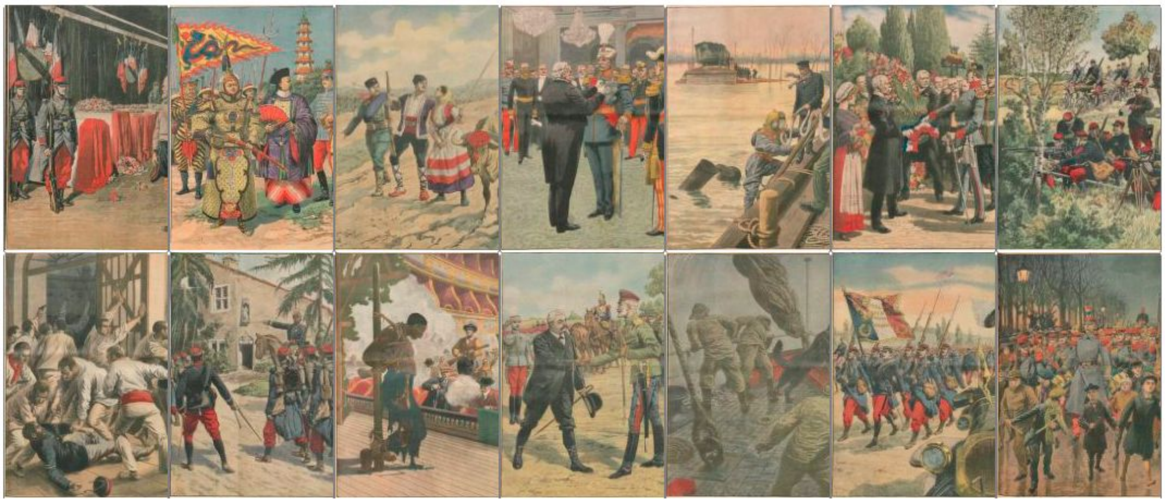
Complex queries: The joint use of metadata and conceptual classes allows the formulation of advanced queries. Figure 24 shows for example the results of a search for visuals relating to the urban destruction following the Battle of Verdun, using the classes “street”, “house” or “ruin”.

Figure 24: Query “class=street AND keyword=Verdun” (samples)



Another example is a study of the evolution of the uniforms of French soldiers during the conflict, based on two queries using the conceptual classes (“soldier”, “officer”, etc.), bibliographic data (“date”), and an image-based criterion (“color”), in order to be able to observe in a couple of mouse clicks the history of the famous red trousers worn until the beginning of 1915.

Figure 25: Query “class=soldier AND mode=color AND date < 31/12/1914”; “date > 01/01/1915” (below)



Other examples of multimodal queries are given in the appendix (Fig. 26 to 28).

5 Further Work

5.1 Experimenting

Several cases of use are being evaluated at the National Library of France: Search of illustrations for digital mediation (see Figure 29 in appendix); Production of ground truths[19] or thematic corpora for research purposes (which still expresses a limited interest [Gunthert17], however growing regarding visual studies which investigates more and more heritage contents [Ginosar15]); Integration of an "image tab" in the Gallica results page. In this latter case, the industrialization of extraction and metadata enrichment processes will be facilitated by the nature of the tasks which tend to be easily parallelizable (at the grain of the illustration or the document). Some future works would also be done regarding the usability and the challenges of searching and browsing into a huge mass of images: clustering, visualization, iterative search driven by user feedback (see [Picard15]), etc.

5.2 Opening the Data

Moving towards sustainability for the metadata describing the illustrations would benefit to their reuse by information systems (e.g. catalogs) as well as by internal softwares used by libraries and also by the users via the data access services (e.g. APIs). The IIIF Presentation API[20] provides an elegant way to describe the illustrations in a document using a "W3C Open Annotation" attached to a layer (Canvas) in the IIIF manifest:

```
{ "@context": "http://iiif.io/api/presentation/2/
context.json",
"@id": "http://example.org/iiif/book1/annotation/anno1",
        "@type": "oa:Annotation",
```

[19] For example, by the mean of bootstrapping with textual descriptors and then a generalization by similarity search.

[20] http://iiif.io/api/presentation/2.1

```
        "motivation": "sc:classifying",
        "resource":{
                "@id": "Ill_0102",
                "@type": "dctypes:Image",
                "label": "photo" },
"on": "http://example.org/iiif/book1/canvas/p1#xywh=30,102,520,308"
}
```

All iconographic resources (identified by manual indexing or OCR) can then be operated by machine, for library-specific projects[21], for data harvesting [Freire17] or for the use of GLAM, hacker/makers and social networks users.

6 Conclusion

Unified access to all illustrations in an encyclopedic digital collection is an innovative service that meets a recognized need. It is part of the effort being made to ensure the greatest value of contents at an appropriate granularity (which implies dropping the "comfortable" page level model and to dig into the digitized contents located in the page) and to open the data in order to promote their reuse. The IIIF protocol can play a major role by allowing to expose and to mutualize these iconographic resources which are increasingly numerous to integrate the patrimonial warehouses.

At the same time, the maturity of modern AI techniques in image processing encourages their integration into the digital library toolbox. Their results, even imperfect, help to make visible and searchable the large quantities of illustrations (which are not manually indexable), of our collections.

We can imagine that the conjunction of this abundance and a favorable technical context will open a new field of investigation for DH researchers in the short term and will offer a new image retrieval service for all other categories of users.

[21] E.g. https://www.flickr.com/photos/britishlibrary/

Appendix

Note: Datasets, scripts and code are available: https://altomator.github.io/Image_Retrieval/

Figure 4: Data Model (XML Schema)

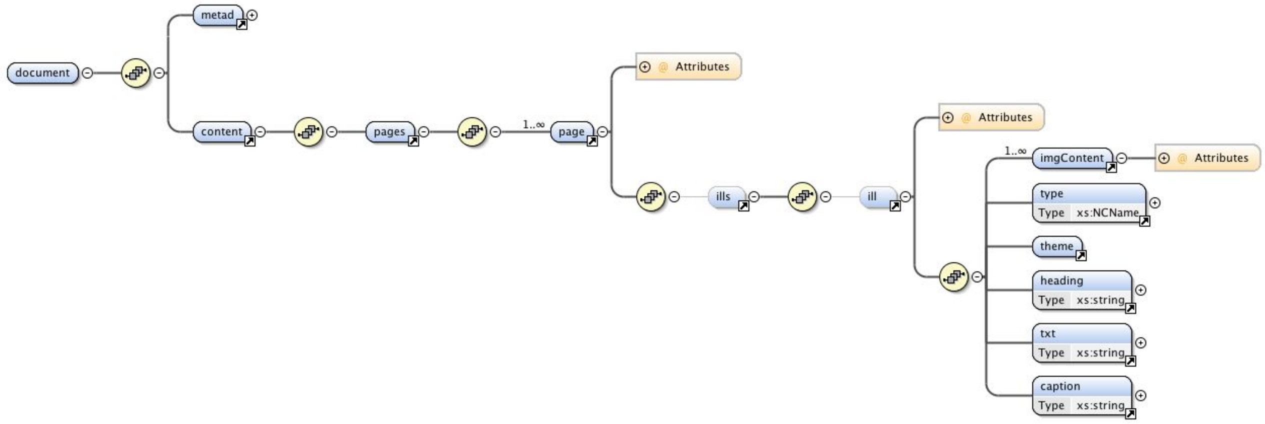


Figure 5: Document sources distribution in the database: on the left, number of pages; on the right, number of illustrations

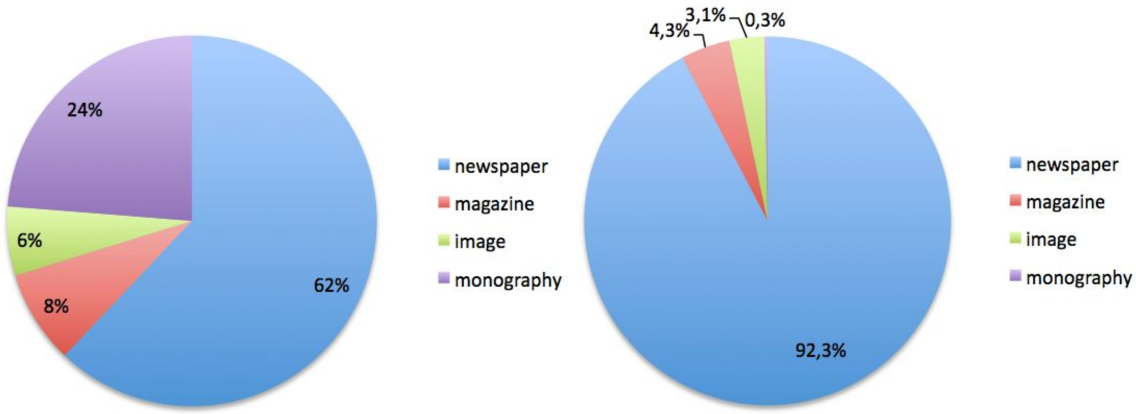


Figure 6: Image corpora

Origine	Contents	Pages	Illustr.
"WW1" OAI set	photo, engraving, map, music score, etc.	9,240	9,240
dc:subject = "WW1"	idem	13,510	13,510
dc:source = "Meurisse"	photo	4,730	4,730
dc:title or dc:subject ⊂ "poster"	poster	610	610

Figure 7: Serial corpora

Type	Title	Pages	Illustr.	After size filtering
Newspapers with article separation (OLR)	*JDPL, Ouest Eclair, Le Gaulois, Le Matin, PJI, Le Parisien, L'œuvre, L'Excelsior*	138,500	164,000	137,290
Newspapers (OCR)	*L'Humanité, Le Figaro, L'Univers, La Croix, La Presse, L'Intransigeant, L'Action, Le Siècle, L'Echo de Paris, Le Constitutionnel, Le Temps*	151,400	661,800	137,000
Sciences magazines	*La Science et la vie, L'Aviation et l'automobilisme militaires, Ligue aéronautique de France, Vie aérienne illustrée, La Restauration maxillo-faciale*	10,500	12,820	12,670
WW1 magazines	*Pages de gloire, Le Miroir, Journal des sciences militaires, L'ambulance, Les Cahiers de la guerre, L'Image de la guerre, La Guerre aérienne illustrée*	27,460	26,240	26,070

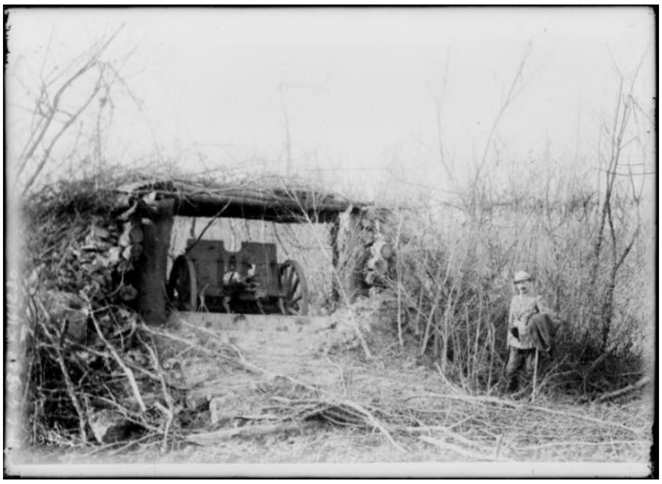
Figure 8: Monography corpora

Type	Nature	Pages	Illustr.	After size filtering
Monographs	History of regiments, misc.	110,870	2,640	2,500

Figure 12: Search results for large illustrations: map, double spread page, poster, comics, etc.

Figure 19: Example of results for a search on musical score covers with red-dominant color

Figure 26: Multimodal query: “class=bunker AND keyword=canon”



This example illustrates an induced advantage of the indexing of image content by a closed vocabulary: independence in the lexicon (or language). The user targeted the “bunker” class and probably would not have thought of the (French) word “casemate” (“blockhouse”), the term used in the bibliographic record that could be described as aged (or technical).

Figure 27: Example of a multimodal query: a wheeled vehicle in a desert environment. The illustration in the middle the is a false positive. (“class=wheeled vehicle AND keyword=sand OR dune”)

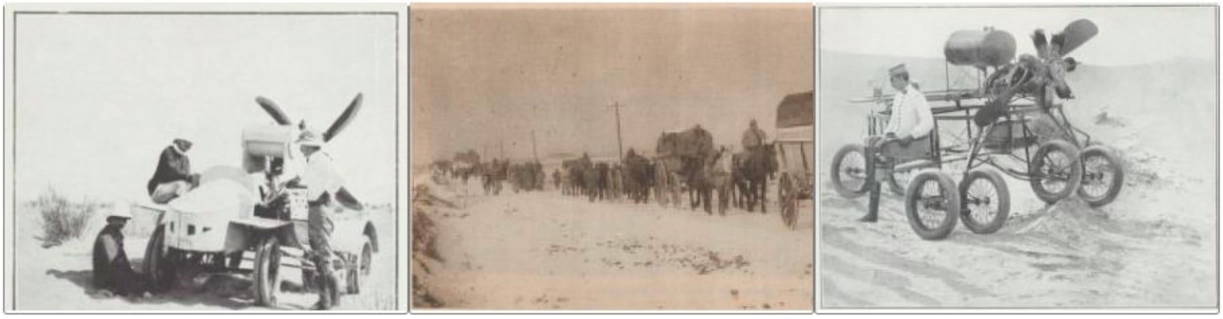
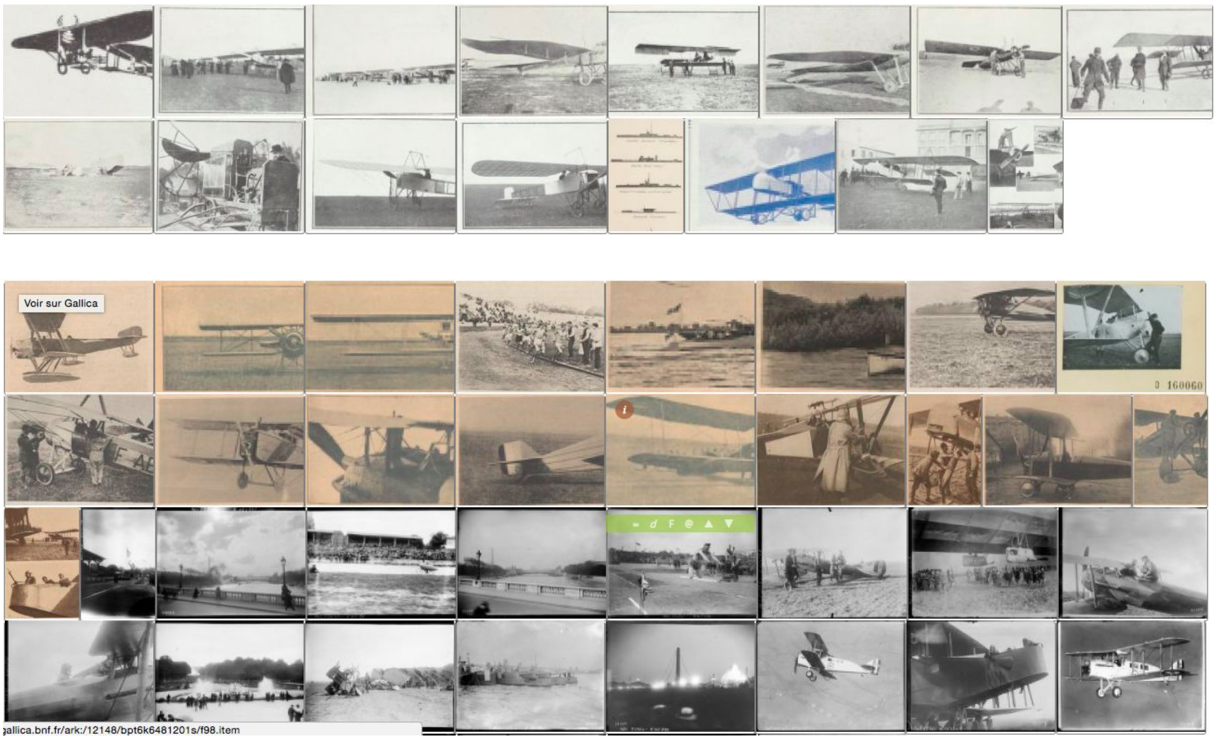


Figure 28 : Example of a multimodal query: history of aviation (samples).
(“class=airplane AND date <= 1914”; “date >= 1918”, bellow)



This last example shows the evolution of aeronautical techniques during the conflict. In this context, the illustrations provided by the system could feed on averaging of images approaches, which increasingly escape the artistic sphere (with human faces as their main subject) to address other subjects (see [Yale14], [Zhu16] et [Feaster16]) or other uses (e.g. automatic dating of photographs, see [Ginosar15]).

Figure 29 : Web portraits gallery based on the results of the face recognition process (see Section 3.4.2)

References

BnF, « Enquête auprès des usagers de la bibliothèque numérique Gallica », avril 2017, http://www.bnf.fr/documents/mettre_en_ligne_patrimoine_enquete.pdf

Breiteneder C., Horst E., “Content-based Image Retrieval in Digital Libraries”, 2000, *Proceedings of Digital Libraries Conference,* Tokyo, Japan, 2000

Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.-P. “Impact of OCR errors on the use of digital libraries”, JCDL'17 ACM/IEEE-CS Joint Conference on Digital Libraries, June 2017, Toronto, Ontario, Canada

Coustaty, M., Pareti, R., Vincent N., Ogier, J.-M., “Towards historical document indexing : extraction of drop cap letters”. *International Journal on Document Analysis and Recognition*, Springer Verlag, 2011, 14 (3), pp.243-254.

Datta R., Joshi D., Li, J., Wang J., “Image Retrieval: Ideas, Influences, and Trends of the New Age”, *ACM Transactions on Computing Surveys*, 2008

Feaster, P., “Time Based Image Averaging”, oct. 2016, https://griffonagedotcom.wordpress.com/2016/10/31/time-based-image-averaging

Freire, N., Robson G., Howard, J.B., Manguinhas H., Isaac, A., “Metadata aggregation: assessing the application of IIIF and Sitemaps within cultural heritage”, TPDL 2017

Ganascia, J.-G., *Le mythe de la Singularité*, Seuil, 2017

Ginosar, S., Rakelly, K., Sachs, S., *et al.*, “A Century of Portraits. A Visual Historical Record of American High School Yearbooks”, *Extreme Imaging Workshop*, International Conference on Computer Vision, 2015, 3

Gordea S., Haskiya D., “Europeana DSI 2– Access to Digital Resources of European Heritage, MS6.1: Advanced image discovery development plan”, http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ Europeana_ DSI-2/Milestones/ms6.1-advanced-image-discovery-development-plan.pdf

Gunthert, A., « Le “visual turn” n’a pas eu lieu », 2017. http://imagesociale.fr/4603

Karpathy A. Fei-Fei, L., “Deep Visual-Semantic Alignments for Generating Image Descriptions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume: 39, Issue: 4, April 1 2017

Lai, H.-P., Visani, M., Boucher, A., Ogier., J.-M., "A new Interactive Semi-Supervised Clustering model for large image database indexing". *Pattern Recognition Letters,* 37 :1–48, July 2013.

Langlais, P.-C., "Identifier les rubriques de presse ancienne avec du topic modeling", 2017, https://numapresse.hypotheses.org

Moreux, J.-P., "Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment Facilitating Access for various Profiles of Users", IFLA News Media Section, Lexington, August 2016

Nottamkandath, A., Oosterman J., Ceolin D., Fokkink W., "Automated Evaluation of Crowdsourced Annotations in the Cultural Heritage Domain", *Proceedings of the 10th International Conference on Uncertainty Reasoning for the Semantic Web*, Volume 1259, Pages 25-36, 2014

Pan, S., Yang, Q., "A survey on transfer learning", *IEEE Transactions on knowledge and data engineering*, volume 22, n. 10, p. 1345-1359, IEEE, 2010

Picard, D., Gosselin, P.-H., Gaspard, M.-C., "Challenged in Content-Based Image Indexing of Cultural Heritage Collections". *IEEE Signal Processing Magazine, Institute of Electrical and Electronics Engineers*, 2015, 32 (4), pp. 95-102

Pourashraf, P., et al., "Genre-based Image Classification Using Ensemble Learning for Online Flyers", Proc. SPIE 9631, Seventh International Conference on Digital Image Processing (ICDIP 2015), 96310Z (July 6, 2015)

Underwood, T., "Topic modeling made just simple enough", 2012*,* https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough

Velcin, J. et al., "Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post", Atelier Journalisme computationnel, Conférence EGC, Grenoble, France, 2017

Wan, G., Liu, Z., "Content-Based Information Retrieval and Digital Libraries", Information Technology and Librairies, March 2008

Wang, K., Q. Yin, W. Wang, S. Wu, and L. Wang. "A comprehensive survey on cross-modal retrieval", 2016. https://arxiv.org/pdf/1607.06215.pdf

Welinder P., Branson, S., Belongie, S., Perona, P. "The Multidimensional Wisdom of Crowds", *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 2424-2432, 2010

Yale University Library, "Robots reading Vogue", 2014, http://dh.library.yale.edu/projects/vogue

Zhu, J.-Y., Lee, Y.-J., Alexei L., Efros, A., "AverageExplorer: Interactive Exploration and Alignment of Visual Data Collections", *ACM Transactions on Graphics* (TOG) 33 (4), 160, 2016