News Media Section

Recherche d'images dans les bibliothèques numériques patrimoniales

Expérimentation à grande échelle des techniques d'apprentissage automatique

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

Résumé : Si historiquement, les bibliothèques numériques patrimoniales furent d'abord alimentées par des images, elles profitèrent rapidement de la technologie OCR pour indexer les collections imprimées afin d'améliorer périmètre et performance du service de recherche d'information offert aux utilisateurs. Mais l'accès aux ressources iconographiques n'a pas connu les mêmes progrès et ces dernières demeurent dans l'ombre : indexation manuelle lacunaire, hétérogène et non viable à grande échelle ; silos documentaires par genre iconographique ; recherche par le contenu (CBIR, *content-based image retrieval*) encore peu opérationnelle sur les collections patrimoniales. Aujourd'hui, il serait pourtant possible de mieux valoriser ces ressources, en particulier en exploitant les énormes volumes d'OCR produits durant les deux dernières décennies (tant comme descripteur textuel que pour l'identification automatique des illustrations imprimées). Et ainsi mettre en valeur ces gravures, dessins, photographies, cartes, etc. pour leur valeur propre mais aussi comme point d'entrée dans les collections, en favorisant découverte et rebond de document en document, de collection à collection. Cet article décrit une approche ETL (*extract-transform-load*) appliquée aux images d'une bibliothèque numérique à vocation encyclopédique : identifier et extraire l'iconographie partout où elle se trouve (dans les collections image mais aussi dans les imprimés : presse, revue, monographie) ; transformer, harmoniser et enrichir ses métadonnées descriptives grâce à des techniques d'apprentissage machine – *machine learning* – pour la classification et l'indexation automatiques ; charger ces données dans une application web dédiée à la recherche iconographique (ou dans d'autres services de la bibliothèque). Approche qualifiée de pragmatique à double titre, puisqu'il s'agit de valoriser des ressources numériques existantes et de mettre à profit des technologies (quasiment) mâtures.

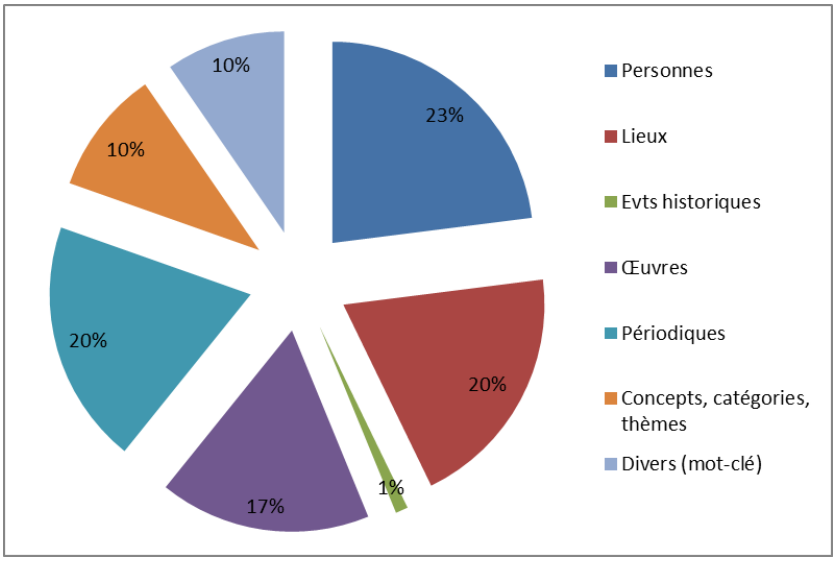
Mots-clés : bibliothèque numérique; recherche d'image; classification d'image ; apprentissage automatique; *machine learning*; *deep learning*; fouille de données; OCR

1 INTRODUCTION

Alors même que la constitution des collections numériques patrimoniales a débuté par l'acquisition en mode image des documents conservés, plusieurs décennies plus tard rechercher dans le contenu de certaines de ces images relève encore d'un futur plus ou moins éloigné [Gordea16]. Ce paradoxe apparent trouve son origine dans deux faits : (1) l'océrisation massive des imprimés a rendu des services majeurs en termes de recherche d'information ; (2) interroger ou parcourir de grandes collections d'images demeure un défi, en dépit des efforts de la communauté scientifique comme des GAFA à relever les défis sous-jacents [Datta08].

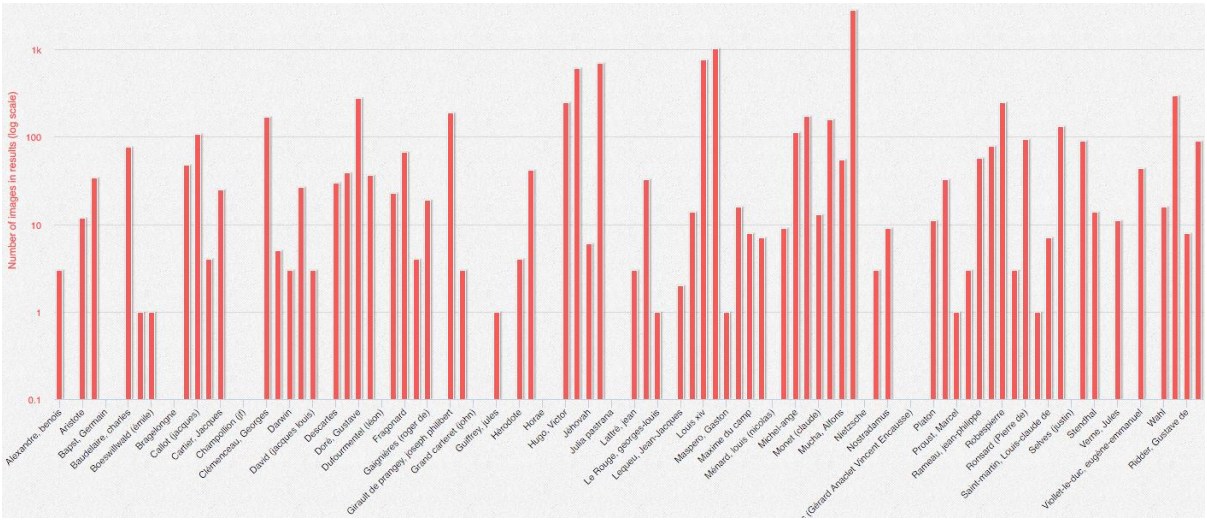
Dans les bibliothèques numériques patrimoniales, les besoins sont pourtant bien réels, si l'on en croit tant les enquêtes auprès des usagers (63 % des utilisateurs de Gallica consultent des images ; 85 % connaissent l'existence d'une collection d'images [BnF17]) que les études statistiques des comportements des usagers : parmi les 500 requêtes les plus courantes, 44 % contiennent des entités nommées Personne, Lieu ou Evénement historique ([Chiron17], cf. figure 1), requêtes pour lesquelles on peut avancer sans risque que des ressources iconographiques apporteraient une information complémentaire à celle présente dans les contenus textuels.

Figure 1 : Répartition des 500 requêtes les plus fréquentes classées par type (gallica.bnf.fr, déc. 2015-mars 2016, 28 millions de requêtes)



Cherchant à répondre à ces requêtes de nature encyclopédique (d'après les chiffres de la figure 1 : 90 % des requêtes ciblent des entités nommées, des titres d'œuvre ou des concepts), les bibliothèques numériques ne sont pas sans ressource. La figure 2 montre ainsi le nombre de résultats donnés (en interrogeant *dc:subject*) par la collection iconographique de Gallica (photographies, gravures, dessins, affiches, cartes, etc.) pour les cent premières requêtes portant sur des entités nommées de type Personne.

Figure 2 : Nombre de documents de Gallica/Images pour les 100 requêtes les plus fréquentes sur une entité nommée de type Personne



Elle met également en lumière des lacunes, liées à sa taille (1,2 M d'illustrations environ), au regard du large spectre des domaines de connaissance et des périodes sondés par les usagers (de l'antiquité au XXI^{e} s.). Pour autant, les bibliothèques sont riches de nombreuses autres sources iconographiques (par ex. la presse, avec jusqu'à trois illustrations par page en moyenne pour les titres les plus illustrés de la première moitié du XX^{e} [Moreux16]). Mais parfois organisées en silos de données peu interopérables, manquant le plus souvent des descripteurs indispensables à la recherche d'image, qu'ils soient « de bas niveau » (taille, couleur, forme) ou sémantiques (indexation thématique, annotation[1], etc.). Et lorsque tous les genres documentaires sont réunis au sein d'un portail unique[2], les modes de recherche et de consultation auront été généralement conçus selon un paradigme classique (métadonnées bibliographiques et indexation en texte intégral). Alors que l'interrogation de fonds iconographiques patrimoniaux pose des enjeux particuliers ([Picard15]), répond à des usages variés (du butinage de photographies anciennes à l'étude des matériaux enluminés ou imprimés de manuscrits ou incunables [Coustaty11]), cible des domaines de connaissance plus ou moins larges (encyclopédique ou spécialisé) et appelle des interactions homme/machine spécifiques (de la saisie d'un mot-clé au tracé d'une ébauche de la forme de l'image recherchée ; voir par ex. [Gang08], [Breiteneder00], [Datta08]).

Cet article présente une proposition de solution pragmatique à ces deux défis, la création d'une base d'images encyclopédique (couvrant plusieurs collections d'une bibliothèque numérique), et celui de ses modalités d'interrogation. Une première section décrit la phase initiale de la nécessaire agrégation des données et métadonnées hétérogènes à disposition (le *e* d'une approche ETL, *extract-transform-load*, schématisée à la figure 3). La deuxième section présente les transformations et enrichissements appliqués aux données collectées, en particulier l'application de traitements relevant des méthodes dites « par apprentissage

[1] Par ailleurs, l'annotation manuelle n'est pas un remède souverain (voir par ex. [Nottamkandath14], [Welinder10]).

[2] C'est le cas de Gallica (http://gallica.bnf.fr).

machine » (ou *machine learning*). Enfin, un mode d'interrogation multimodale est testé et ses résultats sont commentés.

Figure 3 : Processus ETL en trois phases et ses outils

1. Extraction	2. Enrichissement	3. Interrogation
API Gallica	API Watson/Visual Recognition (IBM)	BaseX
OAI-PMH	TensorFlow (Google)	XQuery
SRU	IIIF	IIIF
	Tesseract	Mansory.js
Perl, Python		

2 EXTRAIRE ET AGREGER

Plusieurs décennies après la création des premières bibliothèques numériques patrimoniales (Gallica fête ses vingt ans en 2017), les ressources iconographiques conservées dans les magasins numériques sont à la fois conséquentes et en constante expansion. Des actions de médiation[3,4] et d'indexation manuelle[5] ont été conduites mais leur coût les limite à des corpus restreints et généralement thématiques et/ou monogenres (photographie, affiche, enluminure, etc.). A contrario, une approche massive et multicollection nécessite une première étape d'agrégation de contenus afin de prendre en compte la variabilité des données à disposition, du fait tant de la nature des silos documentaires que de l'histoire des politiques de numérisation ayant présidé à leur constitution.

La base iconographique décrite dans cet article agrège environ 340k illustrations (collectées parmi 465k pages) des collections images et imprimés de Gallica relative à la première guerre mondiale (période 1910-1920). Elle suit un formalisme XML et a été chargée dans une base XML[6] à l'aide des API Gallica[7] et des protocoles SRU et OAI-PMH. Son modèle de données (figure 4, en annexe) agrège les niveaux document, page et illustration ; il permet d'accueillir les informations disponibles dans les différents silos documentaires ciblés (dont la répartition est donnée figure 5, en annexe). L'accès aux illustrations elles-mêmes est réalisé à l'aide du protocole API IIIF Image.

Le retour d'expérience détaillé de cette première étape d'agrégation multicollection est présenté dans les sections suivantes. Notons que cette seule étape justifie l'effort consenti car elle donne aux utilisateurs accès à des illustrations profondément cachées dans les collections numériques. Elle met aussi en lumière certains défis à relever : formats, métadonnées et pratiques de numérisation hétérogènes ; analyse de données massives (mais parallélisable) ; bruit important dans le cas des journaux.

---

[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

2.1 Collections Images

Un *set* thématique préexistant de l'entrepôt OAI-PHM de la bibliothèque numérique est utilisé afin d'en extraire les métadonnées de 6 600 documents images (œuvres graphiques, coupures de presse, médailles, cartes, partitions musicales, etc.) aboutissant à une collection d'environ 9 000 illustrations. Ces documents présentent des défis particuliers : métadonnées souffrant de défauts d'incomplétude et d'inconsistance (du fait de la variabilité des pratiques d'indexation) ; peu ou pas de métadonnées documentaires (genre : photo, gravure, dessin, etc. ; couleur et taille du document originel) ; présence de recueils reliés d'illustrations (les couvertures, pages de texte et pages vierges devant être exclues de la base d'images. Voir figure 9 pour un exemple). Ce corpus (voir figure 6, en annexe) a été complété par diverses requêtes SRU portant sur des métadonnées catalogue (« sujet = Guerre 14-18 », « source = agence photo Meurisse », « type = affiche »).

2.2 Collections Imprimés

La base est alimentée par une sélection intellectuelle d'ouvrages et de revues ainsi que par un échantillonnage temporel de la collection de presse. Ici, le texte océrisé environnant l'illustration est extrait et conservé comme descripteur textuel.

2.2.1 Presse et revue

La collection des périodiques de la BnF se présente sous différents formalismes liés à l'histoire des projets de numérisation successifs. Dans tous les cas, il s'agit d'extraire des formats METS et ALTO[8] les métadonnées descriptives des illustrations. Dans le cas des projets de numérisation récents identifiant les articles (OLR, *optical layout recognition*), cette tâche est facilitée du fait de leur structuration fine et contrôlée ; les plus anciens programmes offrent de l'OCR brut peu structuré. Des revues thématiques enrichissent la base : journaux de tranchées, revues scientifique et technique, revues de sciences militaires, journaux professionnels, etc.

Dans le cas de la presse quotidienne, les illustrations récoltées se caractérisent par des singularités (taille variable des illustrations, de la vignette à la double page ; mauvaise qualité de reproduction, en particulier aux débuts de la photogravure), une grande diversité de genres (de la carte d'état-major à la bande dessinée) et un volume conséquent (figure 7, en annexe). Le bruit est également massif (blocs de texte reconnus erronément par l'OCR comme illustration ; ornements ; publicités illustrées et répétées au fil des publications).

Diverses heuristiques sont appliquées afin de réduire ce bruit, en filtrant sur des critères physiques : taille ; ratio largeur/hauteur (suppression des filets et autres culs-de-lampe) ; emplacement des illustrations (ours de la première page, dernière page d'un fascicule contenant traditionnellement annonces et publicités). Cette étape conduit à 271k illustrations utilisables (sur 826k collectées, soit un bruit de 67%). Un deuxième filtre visant à identifier les publicités et blocs de texte parasites résiduels est réalisé dans une étape ultérieure (cf. section 3.3.1). Notons qu'une recherche par similarité pourrait également permettre de filtrer les publicités et bandeaux de rubriques récurrents.

---

[8]. Le Mechanical Curator de la British Library est une des sources d'inspiration de cet usage exotique de l'OCR (http://mechanicalcurator.tumblr.com).

2.2.2 Monographies

Le même traitement est appliqué à l'OCR des monographies du corpus (figure 10, en annexe) : ouvrages d'histoire, historiques de régiment, etc.

3 TRANSFORMER ET ENRICHIR

Cette étape consiste à transformer, enrichir et aligner les métadonnées obtenues lors de la phase d'agrégation. En effet, les métadonnées descriptives des illustrations récoltées se caractérisent tant par leur hétérogénéité (dans les cas extrêmes, plusieurs centaines d'illustrations sont placées sous une seule notice bibliographique chapeau) que par leur pauvreté au regard des fonctionnalités utilisateur attendues.

3.1 Extraction de texte

Les illustrations des imprimées sans descripteur textuel (par exemple du fait d'une lacune de l'OCR originel) sont détectées et leur emprise englobante est traitée par le moteur OCR Tesseract, ce qui permet d'indexer textuellement les illustrations muettes (voir par ex. la page de une montrée à la figure 23, à droite, qui a été considérée à tort par l'OCR comme une illustration unique).

3.2 Extraction de thèmes

3.2.1 Cas de la collection Images

Un alignement vers l'indexation thématique des contenus de presse IPTC[9] (17 thèmes de premier niveau) est réalisé à l'aide d'une approche par réseau sémantique : les mots-clés des notices des documents (titre, sujet, description) et des légendes des illustrations (quand il y en a) sont lemmatisés puis alignés sur les thèmes IPTC. Une telle méthode n'est pas aisément généralisable (le réseau doit être affiné manuellement en fonction du corpus). Sur un corpus réduit, elle permet cependant d'offrir un classement thématique rudimentaire mais opératoire.

3.2.2 Cas de la collection Imprimés

A contrario, les imprimés se caractérisent par un riche appareil textuel (titraille et légende, texte précédant ou suivant l'illustration) qu'il est possible de thématiser. Les techniques de détection de thèmes seraient[10] ici opérationnelles (cf. par ex. la méthode de *topic modeling* LDA sans apprentissage supervisé [Underwood12], [Langlais17], [Velcin17]). Dans le cas de la presse, média polyphonique par essence, cette thématisation est indispensable. Les corpus de presse numérisés avec une reconnaissance des articles incluent parfois un rubriquage partiel, en général réalisé manuellement par les prestataires de numérisation (petites annonces, publicités, cours de la bourse, chroniques judiciaires, etc.), qu'il est possible d'intégrer dans les métadonnées de classement thématique. Notons que certaines revues thématiques (sciences, sports, arts militaires, etc.) sont également assignables à un thème IPTC.

---

[9] http://cv.iptc.org/newscodes/mediatopic

[10] Tâche non réalisée dans le cadre de cette expérimentation.

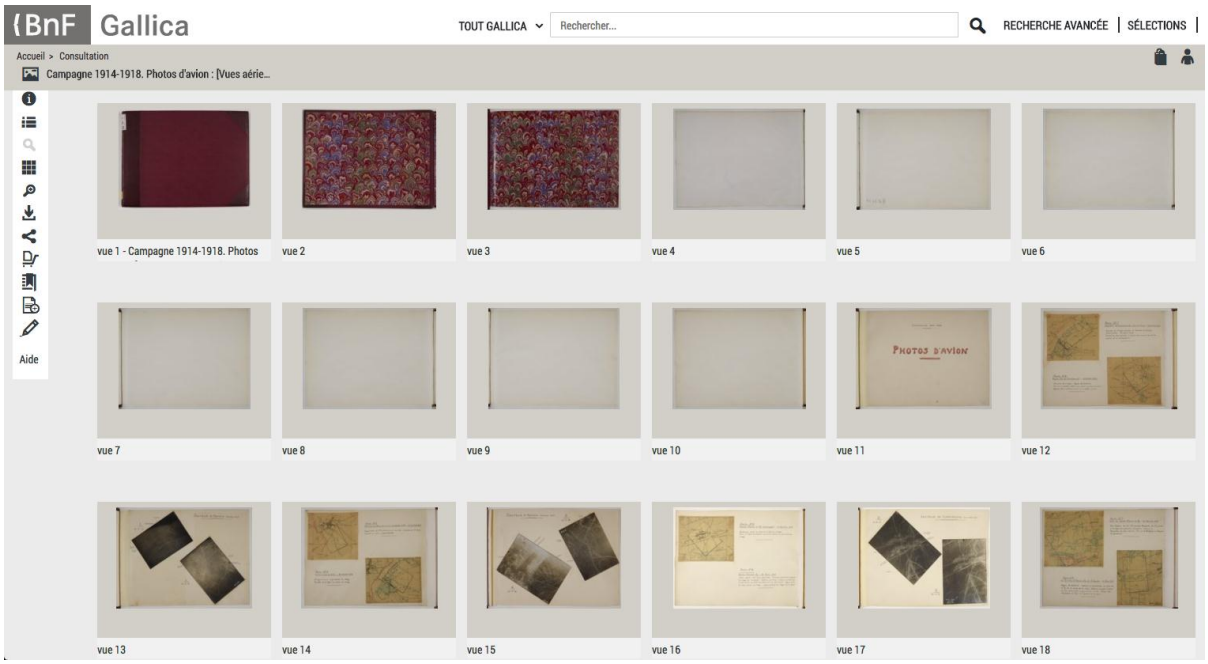
3.3 Extraction de métadonnées image

La recherche dans des contenus image doit affronter une double non-coïncidence : entre la réalité du monde enregistrée dans une scène (dans notre contexte une « illustration ») et la description informationnelle de cette scène ; entre l'interprétation d'une scène par différents utilisateurs (ayant des objectifs de recherche possiblement différents). Réduire ou dépasser ces « fossés » (sensoriel, sémantique) implique notamment de fournir tant aux applications qu'à leurs utilisateurs des descripteurs opératoires (nature des illustrations, couleur, taille, texture, etc.), la recherche opérant ensuite dans l'espace formé par ces descripteurs visuels. La qualité est également un critère à prendre en compte, bien que par nature difficile à quantifier. Pour les photographies patrimoniales par exemple, un distingo est à établir entre tirage argentique et autres modes de reproduction dans les imprimés.

3.3.1 Classification des genres documentaires

Le genre des illustrations (considéré dans une acception large, de la technique de reproduction au type documentaire) n'est pas toujours caractérisé dans les catalogues (figure 9). Cette information n'est bien sûr pas plus disponible pour les illustrations des imprimés.

Figure 9 : Recueil d'illustrations non typées : photo, dessin, carte…



Afin de pallier ce manque, une méthode de classification automatique d'images par apprentissage machine est mise en œuvre.

Les réseaux de neurones modernes (voir [Pourashraf15] pour une approche par SVM) sont capable de reconnaître mille types d'objets de la vie courante (bateau, table, chien…) et surpassent même les humains sur certains jeux de données. Le modèle Inception-v3 [Christian15] est la troisième itération d'amélioration du modèle originel GoogLeNet (un réseau de neurones convolutif à 22 couches ayant gagné la compétition ILSVRC 2014). Les modèles de ce type sont généralement préentraînés sur des supercalculateurs et sont

spécialement optimisés pour exceller sur des bases d'images telles que ImageNet [FeiFei10] (aujourd'hui une référence dans le domaine tant pour sa taille que pour sa représentativité). L'effort toujours plus important apporté à l'amélioration de ces modèles bénéficie à la communauté « vision par ordinateur » en général mais pas seulement. En effet, il est aujourd'hui possible d'exploiter la puissance capitalisée par ces modèles sur des problématiques autres (ici la classification de documents patrimoniaux). Cela nécessite un réapprentissage « léger » d'une partie du modèle (en fait sa dernière couche, ce qui nécessite quelques heures sur une machine classique, à comparer aux semaines ou mois qui seraient nécessaires pour réentraîner le modèle complet), en suivant une approche dite de « *transfer learning* » [Pan09]. Cette approche consiste à réutiliser les descripteurs visuels élémentaires trouvés durant la phase d'apprentissage originelle (dans la mesure où ils ont prouvé leur capacité à classifier un jeu de données), mais sur un nouveau jeu de données avec l'espoir que ces descripteurs continueront à bien se comporter. De plus, en réduisant le nombre de classes (ce qui simplifie le problème en quelque sorte), il est tout à fait possible de conserver des scores honorables de classification, alors même que l'on utilise un modèle n'étant pas spécifiquement entraîné pour la tâche en question.

La figure 10 donne un aperçu des douze catégories de documents que l'on cherche à faire apprendre au modèle Inception-v3 suivant l'approche décrite (bande dessinée, carte, dessin, gravure, écriture manuscrite, partition, photo, publicité, texte, page blanche, couverture, ornement).

Figure 10 : Les classes constituant le jeu de données d'apprentissage
(*nombre de documents entre parenthèses)*

Cet apprentissage nécessite de fournir un certain nombre de documents étiquetés par leur classe (7786 dans notre exemple). Une fois entraîné, le modèle est ensuite évalué à l'aide

d'une base de test (1952 documents). La figure 11 détaille les résultats obtenus. Le rappel moyen[11] est de 0,90 et la précision[12] est de 0,90, ce qui correspond à un F-mesure[13] de 0.90. On peut considérer ces résultats comme bons au regard du rapport taille/diversité du jeu de données utilisé pour l'apprentissage. Notons que les performances sont meilleures avec un modèle moins générique (entrainé sur le seul corpus des monographies, le F-mesure monte à 0.94). Abandonner l'approche par *transfert learning* pour un modèle totalement entraîné (mais au prix d'un temps de calcul très supérieur) aurait également un effet bénéfique.

Figure 11 : Résultats de reconnaissance sur les douze classes du modèle

Documents belonging to ↓ / Recognized as →	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	

Ce modèle est également utilisé pour filtrer les types documentaires indésirables, en particulier les blocs de texte parasites de la presse (et éventuellement les publicités illustrées) et les couvertures et pages blanches des recueils d'images (cf. section 2.1). L'application de ce filtre sur les 6 000 illustrations d'un titre de presse complet[14] sans iconographie exploitable (exception faite des publicités) conduit à un taux de rappel de 0,983 (plus de 98% du bruit est supprimé).

3.3.2 Taille, couleur, localité

Quand elle n'est pas disponible dans les métadonnées de numérisation, le mode colorimétrique de chaque illustration est extrait. Un cas particulier concerne les documents originellement monochromatiques (noir et blanc, sépia, sélénium, etc.) numérisés en couleurs, pour lesquels une approche naïve d'analyse de la composante *hue* du modèle HSV est utilisée (voir aussi section 3.4.3).

La localité, la taille et la densité des illustrations sont également extraites. Dans le cas de presse, interroger la seule une ou rechercher une illustration de grande taille (figure 12, en annexe) sont représentatifs de cas d'usage courants et légitimes.

---

[11] Nombre de documents pertinents retrouvés par le classifieur au regard du nombre de documents pertinents que possède la base.

[12] Nombre de documents pertinents retrouvés au regard du nombre total de documents retrouvés par le classifieur.

[13] Mesure qui combine la précision et le rappel.

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

3.4 Extraction des contenus image

Historiquement (voir [Datta08]), tout système CBIR se devait d'extraire les descripteurs visuels d'une image, d'en déduire une signature puis d'opérer la recherche dans l'espace des signatures à l'aide d'une mesure de similarité, ce qui implique de fournir la requête sous la forme d'une signature (c'est-à-dire une image, cette contrainte ayant un impact négatif sur l'utilisabilité de ces systèmes [Gang08]). De plus, il est apparu qu'une mesure de similarité peinait à transcrire la richesse sémantique et la subjectivité d'interprétation des contenus image, en dépit des améliorations apportées (par ex. la prise en compte des sous-régions d'une image).

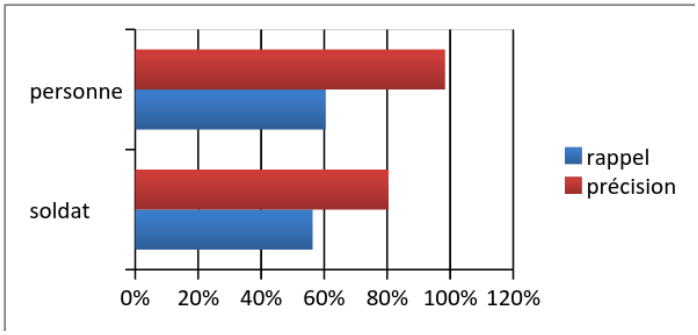
Plus récemment, les progrès des techniques d'apprentissage machine ont permis de dépasser ces limites, en particulier grâce à des approches dites de *clustering* et de classification (ou extraction de concepts), cette dernière ayant en outre l'avantage de formuler la requête sous forme textuelle [Karpathy17]. L'API Visual Recognition d'IBM (qui s'inscrit dans les services rendus par l'IA Watson[15]) est exemplaire de ces évolutions[16] et les sections suivantes décrivent son application à une base d'images patrimoniales.

3.4.1 Détection de concepts

Visual Recognition applique des algorithmes d'apprentissage profond à l'analyse d'images afin d'en extraire des concepts (objets, personnes, couleurs, etc.) identifiés au sein de la taxonomie des classes connues de l'API. Elle renvoie des paires classe/estimation de confiance.

Une évaluation est menée sur la détection de personnes. Une vérité terrain de 2 200 images est créée (VT), couvrant la variété documentaire de la base (photo, gravure, dessin) et respectant une répartition de 80/20 entre illustrations avec et sans personne représentée (qui est également représentative de la collection). Une autre évaluation est menée sur la classe Soldat (600 images, avec une répartition 50/50).

Figure 13 : Rappel et précision des classes Personne et Soldat



Pour la classe Personne, le taux rappel de 60,5 %[17] conduit à fournir 1 190 illustrations à l'utilisateur avec une excellente précision (98,4 %). Les taux sont moindres sur une classe

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

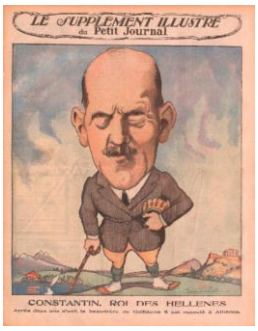
[16] L'API TensorFlow Object Detection (Google) aurait également pu être utilisée.

[17] Une VT moins « stricte » (toute illustration incluant une silhouette, même de petite taille, a été comptée dans la VT) conduirait à un meilleur rappel.

plus spécialisée (Soldat, rappel : 56 %, précision : 80,5 %), mais ils sont à mettre en regard du relatif silence des approches classiques : le concept “personne” n’existe pas dans les métadonnées bibliographiques (a fortiori pour des illustrations d’imprimés non cataloguées), et une recherche sur le mot-clé “personne” dans la VT ne renvoie que 11 illustrations correctes. Par le même moyen, “soldat” renvoie 59 résultats et il faudrait écrire une requête complexe du type “soldat OU officier OU militaire OU artilleur OU poilu ...” pour aboutir à un taux de rappel de 21 %, à mettre en regard des 56 % obtenus avec la reconnaissance visuelle. Il convient de noter un appréciable taux de rappel combiné (utilisation simultanée des descripteurs textuel et visuel) de 70 %.

Le rappel de la classe Personne par genre est également analysé : gravure et dessin : 54 % ; photo argentique : 67 %, photogravure : 72 %. L’API montre sa capacité à traiter des documents « difficiles » (voir figure 14).

Figure 14 : Exemples de résultat pour la classe Personne


Le type photogravure présente un taux supérieur à celui de la photo argentique, ce qui peut sembler surprenant, mais s’explique par la complexité des scènes de la collection d’images comparée à celle des illustrations des imprimées (scènes plus simples, de format inférieur). De manière générale, les scènes complexes (multiobjets) mettent en évidence les limites actuelles de ces technologies et la nécessité de les dépasser (voir par ex. un modèle génératif de description en langage naturel d’images et de leur région [Karpathy17]). La figure 16 montre des exemples de ce type, ainsi qu’un autre cas malheureux, celui des documents encadrés, qui sont classées en tant que tel (et non d’après leur contenu).

Figure 15 : Scène complexe (g.) : l’API suggère « explosive device » et « car bomb » mais ne détecte pas de personne ; illustration encadrée (d.)


3.4.2 Détection de visages

Les *gender studies* forment un champ de recherche à part entière et le réemploi de visuels numériques de visages humains pour des activités récréatives [Feaster16] ou scientifiques [Ginosar15] a ses praticiens. Il n'est donc pas anodin pour une bibliothèque numérique de prendre en compte de tels besoins.

L'API Watson offre un service de détection de visages, qui fournit en outre les âge et genre (H/F) estimés des personnes. Les résultats sur la VT sont présentés figure 16. On peut constater un taux de rappel de 30 % pour les visages (H+F) et une précision proche de 100 %. Il s'agit d'un corpus difficile pour ce genre d'exercice, car incluant dessin, gravure, photos dégradées, etc. (cf. figure 17). Le taux de rappel est de 22 % pour les classes H et F et la mauvaise précision de 26,5 % pour la classe Femme (l'API a tendance à peupler le monde de femmes à moustaches...). En imposant un seuil de 50 % à l'estimation de confiance (fournie par l'API), la précision pour la classe F s'améliore mais au détriment du rappel.

Figure 16 : Rappel et précision des classes Visage, Homme, Femme

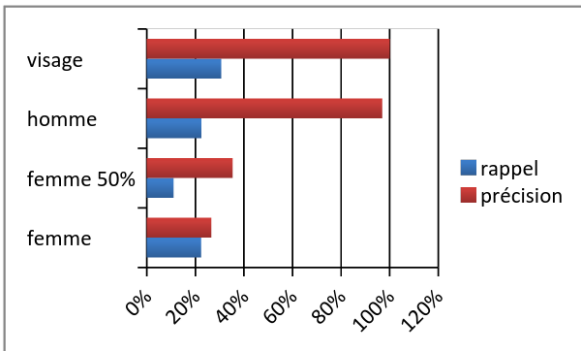


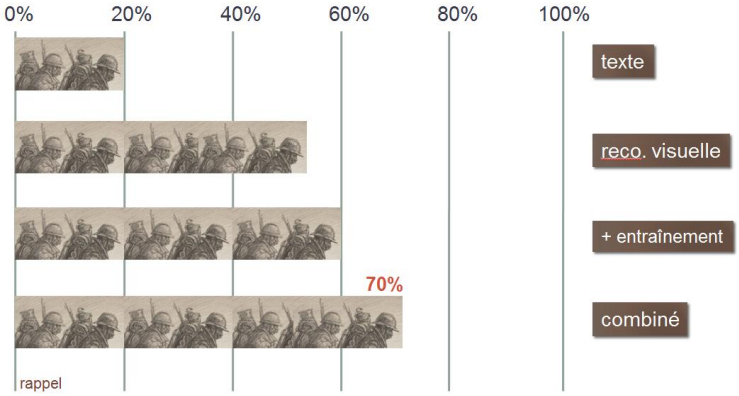
Figure 17 : Exemples de résultats pour la détection des visages

Notons que l'utilisation cumulée des résultats des deux API de reconnaissance (classe Personne et détection de visages) aboutit à améliorer le taux de rappel global pour la détection de personnes à 65 %.

L'API permet également de créer des classifieurs entraînés sur des données locales. Cette possibilité est mise à profit en définissant un corpus d'apprentissage Personne(Homme, Femme)/Non_Personne. La VT est ensuite réanalysée. Il est constaté une amélioration significative de la détection des personnes (rappel : 65 %, précision : 93 %) mais aussi peu d'effet sur la détection du genre. Le rappel global (en utilisant conjointement la classe générique de l'API et le classifieur local) est également amélioré (85 %).

La figure 18 résume les taux de rappel pour la classe Soldat selon les quatre modalités d'interrogation étudiées (descripteurs textuels ; reconnaissance visuelle ; reconnaissance visuelle avec classifieur local ; combiné : texte+visuel) et montre l'intérêt évident à offrir aux utilisateurs une recherche multimodale.

Figure 18 : Taux de rappel de la classe Soldat pour quatre modalités de recherche



3.4.3 Détection des couleurs

Les classes de couleur fournies par l'API (une à deux couleurs dominantes par image) peuvent être mises à disposition des utilisateurs afin d'interroger la base d'images sur ce critère (figure 19, en annexe).

Notons enfin que l'API Watson propose également une fonctionnalité de recherche d'image par similarité[18] (non évaluée ici).

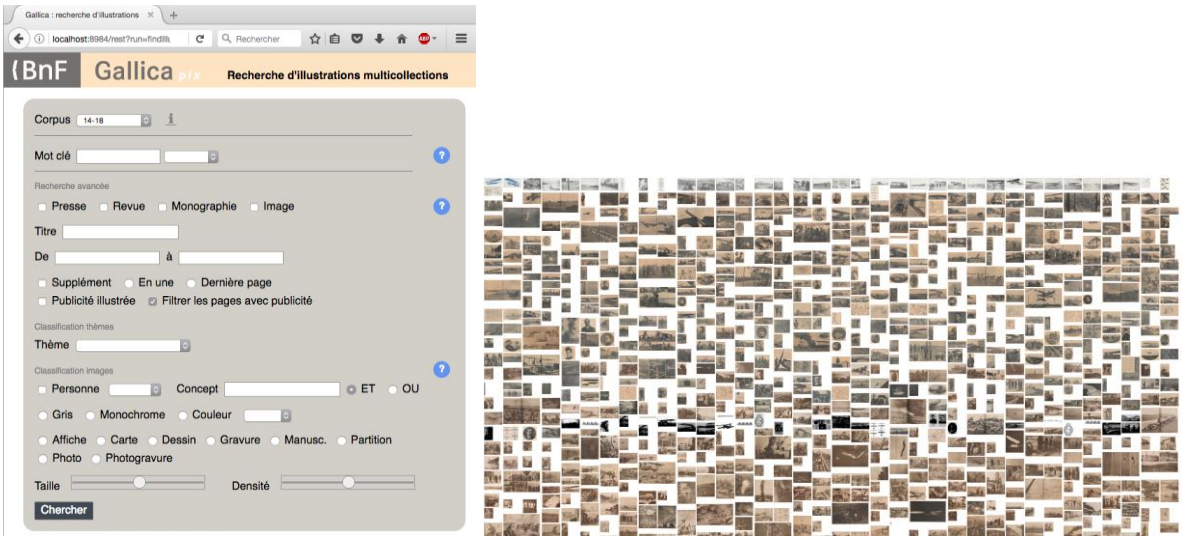
4 CHARGER ET INTERROGER

Les métadonnées XML sont chargées dans BaseX et interrogées en mode client/serveur REST avec un formulaire HTML et des expressions XQuery/FLOWR. La mosaïque d'images est créée avec la bibliothèque Javascript Mansory et peuplée par le serveur IIIF de Gallica. Un système rudimentaire de facettes (couleur, taille, type, date, etc.) permet de préfigurer ce que serait une interaction utilisateur/système aboutie.

[18] Voir par exemple https://bildsuche.digitale-sammlungen.de pour une implémentation à grande échelle de recherche par similarité.

La complexité du formulaire de recherche et le grand nombre de résultats qu'il fournit souvent (voir figure 20) rappelle, s'il en était besoin, que chercher et naviguer dans des bases d'images de grande dimension pose des problèmes spécifiques d'usage et constitue un sujet de recherche à part entière (voir par exemple [Lai13]). Ainsi les modalités opérationnelles de l'interrogation multimodale (au sens de [Wang16] : par les contenus image et par les descripteurs textuels) doivent être rendues intelligibles aux utilisateurs. De même la présence de faux positifs et de bruit dans les résultats fournis (mais ce paysage est proche de celui de l'OCR, dont sont désormais familiers les usagers des bibliothèques numériques patrimoniales). Cela étant, ce mode de recherche contribue à réduire l'écart entre la formulation du besoin utilisateur (par ex. « un visuel de bonne qualité d'une classe d'école en 1914 ») et le modèle de données qui est intelligible au système d'information, comme l'illustrent les exemples suivants. Ils détaillent des cas d'usage représentatifs de l'interrogation d'une base d'images patrimoniales à vocation encyclopédique.

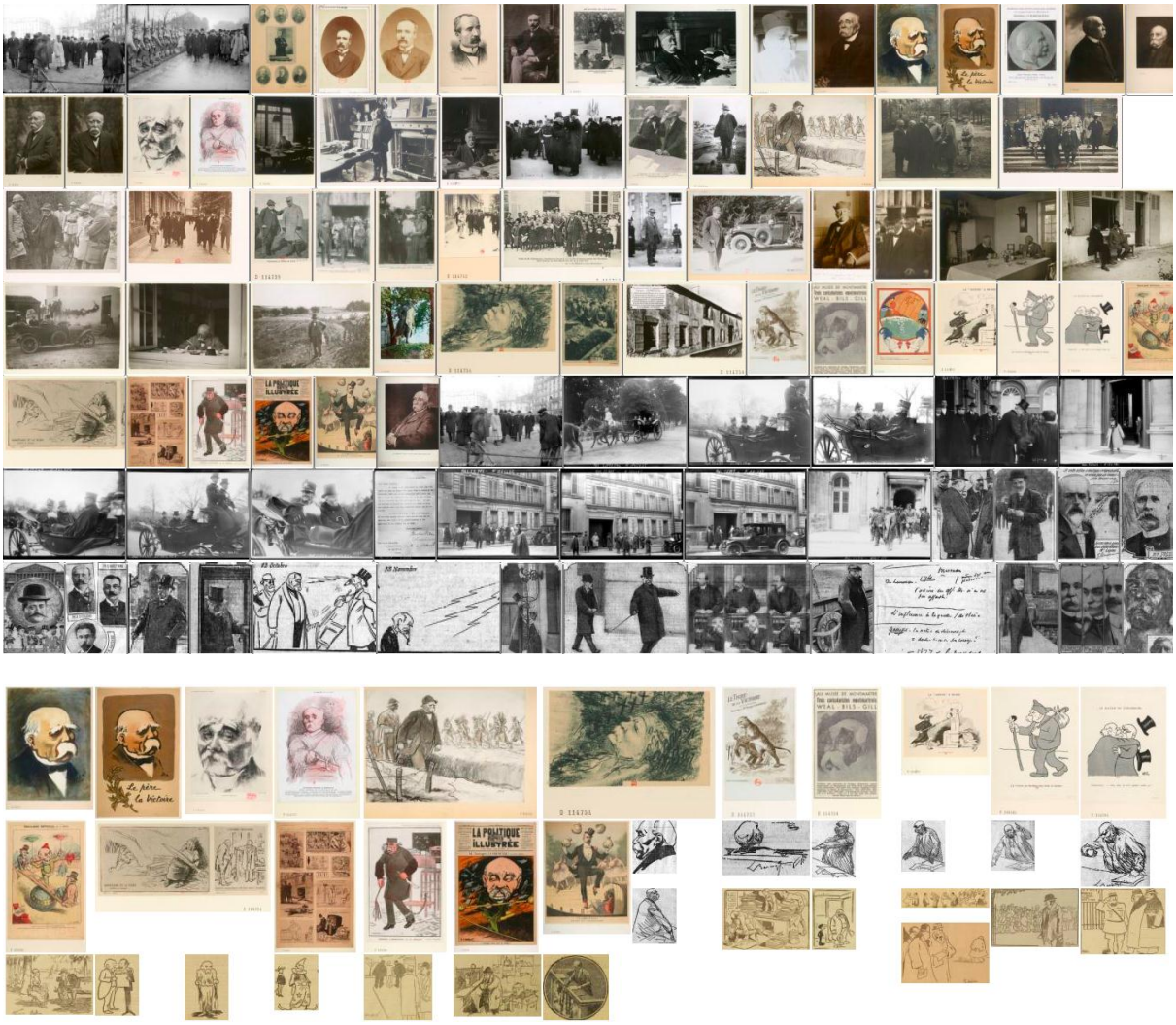
Figure 20 : Le formulaire de recherche (g.) et une représentation de la cardinalité des résultats (d.)



Requête encyclopédique sur une entité nommée : les descripteurs textuels (métadonnées et OCR) sont mis à contribution. Une des cent premières requêtes sur une entité nommée de type Personne posée par les utilisateurs de Gallica porte sur Georges Clemenceau (cf. section 1, avec 130 documents résultats sur la période 1910-1920). La même requête renvoie désormais plus de 1 000 illustrations d'un large spectre de genres.

Les facettes peuvent ensuite être mises en œuvre pour affiner la recherche (par exemple en filtrant les dessins et autres caricatures de presse, voir figure 21, en bas). Précision et rappel sont alors liés à la qualité des descripteurs textuels.

Figure 21 : Requête : "Clemenceau" (extraits) ; en bas, avec la facette « dessin »

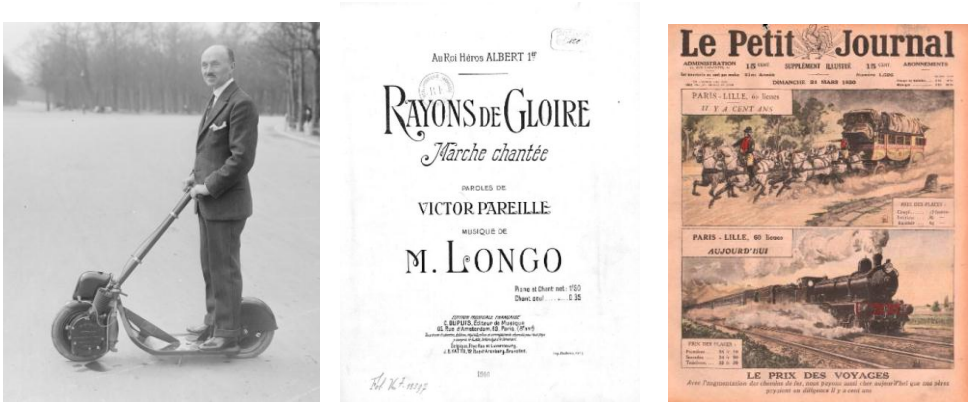


Requête encyclopédique sur un concept : les classes conceptuelles extraites par l'API Watson permettent de lever le silence des métadonnées bibliographiques ou celui de l'OCR, mais aussi de contourner les difficultés liées aux corpus multilingues (certains documents de la base d'images sont d'origine allemande) ou à l'évolution des lexiques (cf. figure 26 en annexe). Les thématiques IPTC (ou tout autre système d'indexation de contenus) ressortent du même cas d'usage. Dans le contexte de la Grande Guerre, on pense ici par exemple aux personnes (cf. supra : genre, soldat), véhicules, armes, etc. Dans ce cas d'usage, l'utilisateur n'attend généralement pas exhaustivité et précision mais plutôt des propositions. La figure suivante montre l'exemple d'une requête sur la superclasse « véhicule », qui fournit de très nombreuses instances de ses sous-classes (vélo, avion, bateau, ballon dirigeable, etc.).

Figure 22 : Requête : classe="véhicule" (extraits)

Les effets de l'apprentissage machine se font parfois sentir, et en particulier ceux liés au processus sous-jacent de généralisation. Ainsi cette trottinette à moteur de 1917 est-elle étiquetée « Segway » (figure 23, à gauche) ou cette page de titre d'une partition musicale (au milieu) indexée comme étiquette de grand cru de bourgogne. Et une locomotive à vapeur (à droite) revient à la lumière sous la dénomination de « véhicule blindé ». Gardons à l'esprit que ces techniques restent tributaires des modalités de création des corpus d'apprentissage [Ganascia17] et que mêmes les plus avancées d'entre elles, entraînées à reconnaître des chiens jouant au frisbee [Karpathy17], ne seront pas à leur avantage sur des documents du début du XX^{e} siècle...

Figure 23 : Biais de l'apprentissage machine



Requête multimodale : l'utilisation conjointe des métadonnées et des classes conceptuelles autorise l'expression de requêtes avancées. La figure 24 montre par exemple les résultats d'une recherche de visuels se rapportant aux destructions urbaines consécutives à la bataille de Verdun, en utilisant les classes « rue », « maison » ou encore « ruine ».

Figure 24 : Requête : classe="rue" ET mot-clé="Verdun" (extraits)

Autre exemple, une étude de l'évolution des uniformes des soldats français au cours du conflit pourra s'appuyer sur deux requêtes mettant en œuvre les classes conceptuelles (« soldat », « officier », etc.), les données bibliographiques (date) et un critère image (« couleur »), afin de mettre en évidence en quelques clics l'histoire du célèbre pantalon garance porté jusqu'au début de l'année 1915.

Figure 25 : Requête : classe="soldat" ET mode="couleur" ET date avant "31/12/1914" ; date après "01/01/1915" (en dessous)

D'autres exemples de requêtes multimodales sont donnes en annexe (figures 26 à 28).

5 FUTURS TRAVAUX

5.1 Expérimenter

Plusieurs cas d'usage sont en cours d'évaluation à la BnF : recherche d'illustrations pour la médiation numérique (voir figure 29 en annexe) ; production de vérités terrains[19] ou de corpus thématiques pour la recherche (qui exprime un intérêt encore limité [Gunthert17], mais cependant croissant, pour les études visuelles ayant pour terrain d'investigation des contenus patrimoniaux, voir par ex. [Ginosar15]) ; intégration d'un onglet « Images » dans la page de résultats de Gallica. Dans ce dernier cas, l'industrialisation du processus d'extraction et d'enrichissement des métadonnées sera facilitée par la nature des traitements, aisément distribuables et parallélisables (au grain de l'illustration ou du document). De futurs travaux devront aussi être consacrés aux défis d'utilisabilité posés par la navigation et la recherche dans une grande masse d'images : clusterisation, visualisation, recherche itérative pilotée par les retours de l'utilisateur (voir par ex. [Picard15]), etc.

5.2 Diffuser

Les métadonnées descriptives des illustrations gagneraient à être pérennisées afin de favoriser leur réutilisation tant par les systèmes d'information (par ex. les catalogues) et applicatifs internes de la bibliothèque que par les usagers, via les services d'accès aux données. L'API IIIF Presentation[20] offre un moyen élégant de décrire dans le manifeste IIIF les illustrations présentes dans un document, sous la forme d'une liste d'annotations (W3C Open Annotation) attachée à un calque (*Canvas*) :

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

La totalité des ressources iconographiques (identifiées par indexation manuelle ou par l'OCR) devient alors actionnable par machine, pour des projets propres à la bibliothèque[21], pour le moissonnage de données [Freire17] ou à l'usage des communautés GLAM, hackers/makers et des utilisateurs des réseaux sociaux.

[19] Par exemple à l'aide d'un amorçage avec les descripteurs textuels puis d'une généralisation par similarité.

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

[21] Voir par exemple https://www.flickr.com/photos/britishlibrary

6 CONCLUSION

L'accès unifié à toutes les illustrations d'une collection numérique encyclopédique est un service innovant répondant à un besoin attesté. Il participe de l'effort en cours de valorisation des contenus à la granularité adéquate (ce qui implique d'abandonner le confortable modèle de la page numérisée et d'entrer dans cette dernière) et d'ouverture des données afin de favoriser leur réutilisation. Le protocole IIIF peut ici jouer un rôle majeur, en permettant d'exposer et de mutualiser les ressources iconographiques, toujours plus nombreuses à intégrer les entrepôts patrimoniaux.

Dans le même temps, la maturité des techniques d'IA en matière de traitement d'images encourage à leur intégration dans la boîte à outils des bibliothèques numériques. Leurs résultats, mêmes imparfaits, contribuent à rendre visibles les riches ressources iconographiques de nos collections — dont l'indexation manuelle est hors de portée.

On peut imaginer que la conjonction de cette abondance et d'un contexte technique favorable permettra d'ouvrir à court terme un nouveau champ d'investigation pour les chercheurs et de nouveaux services de recherche iconographique pour tous les usagers.

Annexes

Note : Jeux de données, scripts et code sont disponibles :
https://github.com/altomator/Image_Retrieval

Figure 4 : Modèle de données

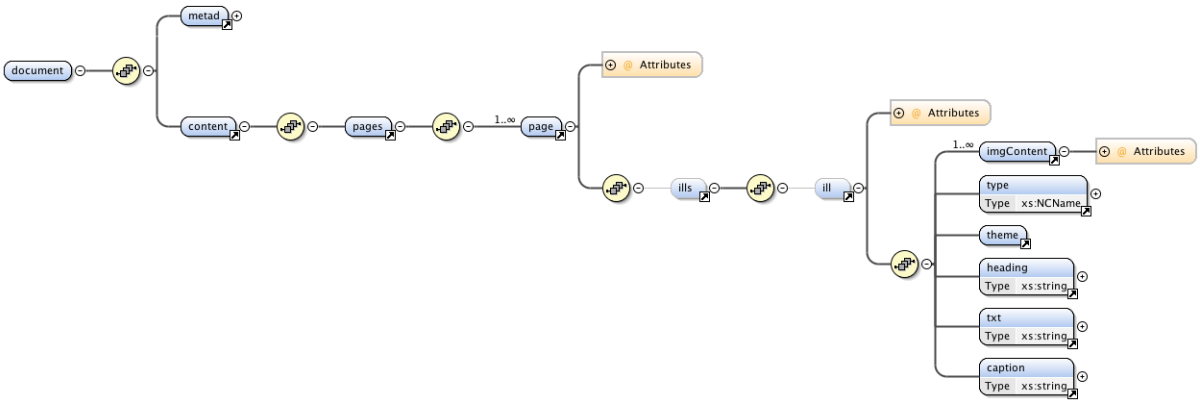


Figure 5 : Répartition des sources documentaires dans la base d'images : à gauche, en nombre de pages ; à droite en nombre d'illustrations

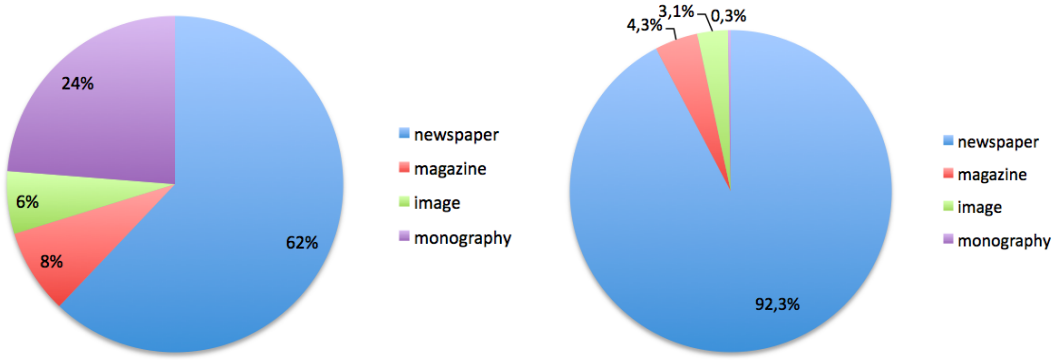


Figure 6 : Corpus « Image »

Origine	Contenus	Pages	Illustr.
set OAI « 14-18 »	photo, gravure, carte, partition, etc.	9 240	9 240
dc:sujet = « 14-18 »	idem	13 510	1 3510
dc:source = « meurisse »	photo	4 730	4 730
dc:titre ou dc:sujet ⊂ « affiche »	affiche	610	610

Figure 7 : Corpus « Presse et revue »

Type	Titre	Pages	Illustr.	Après filtrage taille
Presse quotidienne avec reconnaissance des articles (OLR)	*JDPL, Ouest Eclair, Le Gaulois, Le Matin, PJI, Le Parisien, L'Œuvre, L'Excelsior*	138 500	164 000	137 300
Presse quotidienne (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
Revues sciences et techniques	*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
Revues « 14-18 »	*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 : Corpus « Monographie »

Type	Nature	Pages	Illustr.	Après filtrage taille
Monographies	Historiques de régiment, divers	110 870	2 640	2 500

Figure 12 : Recherche d'illustrations de grande taille : carte, illustration en double page, affiche, bande dessinée, etc. (extraits)

Figure 19 : Recherche de partitions musicales « patriotiques »
(avec une couleur dominante rouge)

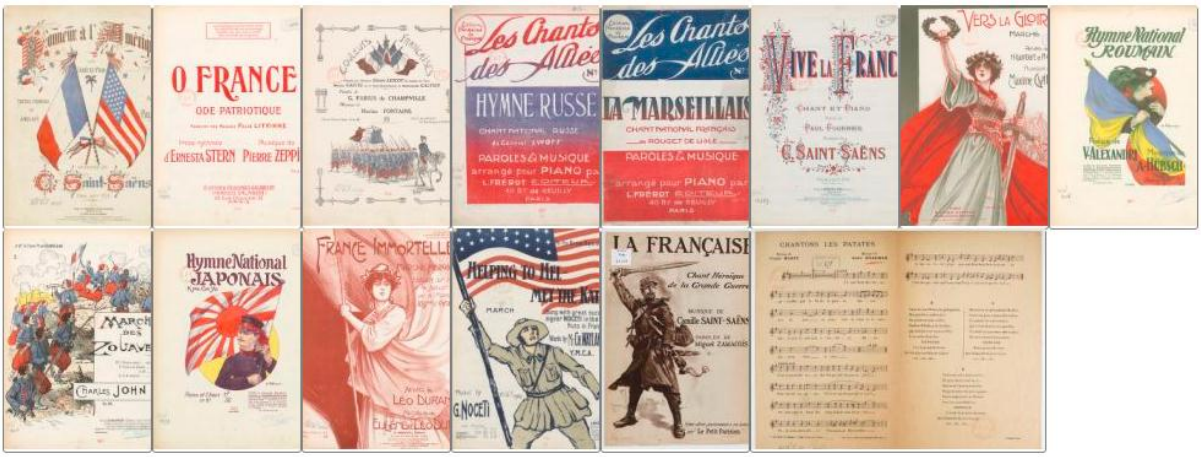


Figure 26 : Exemple de requête multimodale : recherche d'un visuel de canon dans un bunker
(classe="bunker" ET mot-clé="canon")

Cet exemple illustre un avantage induit de l'indexation des contenus image par un vocabulaire fermé : l'indépendance au lexique (ou à la langue). L'utilisateur a employé la classe « bunker » et n'aurait probablement pas pensé au mot-clé « casemate », terme de la notice bibliographique de l'image que l'on pourrait qualifier de vieilli (ou technique).

Figure 27 : Exemple de requête multimodale : véhicule à roues dans un environnement désertique. L'illustration du milieu est un faux positif.
(classe="wheeled vehicle" ET mot-clé="sable" OU "dune")

Figure 28 : Exemple de requête multimodale : histoire de l'aviation.
(classe="airplane" ET date <= "1914", en haut ; date >= "1918", en bas)

Ce dernier exemple donne à voir l'évolution des techniques aéronautiques durant le conflit. Dans ce contexte, les illustrations fournies par le système pourraient alimenter des techniques de moyennage d'images, lesquelles échappent de plus en plus à la seule sphère artistique (avec pour principale matière les visages humains) pour aborder d'autres sujets (voir [Yale14], [Zhu16] et [Feaster16]) ou d'autres usages (par ex. la datation automatique de photographies, voir [Ginosar15]).

Figure 29 : Galerie de portraits générée d'après les données de la détection automatique de visages (voir section 3.4.2)

Références

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 librairies", 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