

# Étude sur l'utilisation d'œuvres relevant des arts visuels dans les publications scientifiques

Pierre-Carl Langlais, Antoine Blanchard

▶ **To cite this version:**

Pierre-Carl Langlais, Antoine Blanchard. Étude sur l'utilisation d'œuvres relevant des arts visuels dans les publications scientifiques. [Rapport de recherche] Ministère Enseignement supérieur et recherche. 2022, pp.88. hal-03682113

**HAL Id: hal-03682113**
**https://hal.science/hal-03682113**

Submitted on 30 May 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

*Liberté*
*Égalité*
*Fraternité*

# Étude sur l'utilisation d'œuvres relevant des arts visuels dans les publications scientifiques

Étude commandée par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation

**Auteurs** : Pierre-Carl Langlais et Antoine Blanchard

**Date** : Août 2021

DATACTIVIST

Ce rapport est placé sous licence Creative Commons CC-BY, à l'exclusion des contenus protégés par le droit des tiers, notamment les images issues du corpus.

**Contexte et structure de l'étude** 3

**Phase 1 : délimitation du corpus et dénombrement des images** 6

**Extraction des documents** 6

La récupération des métadonnées sur Isidore 6

Délimitation du corpus de documents 7

**Extraction des images** 12

Corpus HTML 12

Corpus PDF 13

**Classification des images** 15

Catégories du modèle 15

Évaluation du modèle 18

Résultats de la classification 18

**Métadonnées des images** 25

Légendes 25

Métadonnées internes 25

Estimation du nombre d'images caviardées 27

**Phase 2 : analyse documentaire d'un échantillon** 28

Objectifs 28

Construction de l'échantillon 28

**Méthode d'analyse** 30

Champs de description 31

Sources de référence 35

**Résultats de l'analyse** 36

Résultat général 36

Résultats complémentaires 44

Commentaires 47

**Phase 3 : test d'une méthode automatique reproduisant les résultats de l'analyse manuelle sur l'échantillon** 48
Objectifs 48
Échantillon test 49
Classification des images 49
Analyses automatisée des légendes 51
Extraction des légendes des documents au format PDF 51
Classification des légendes 53
Extraction des entités nommées 55
Identification des reprises d'images 60
Vers un workflow automatique 62

**Phase 4 : estimation du nombre d'images dans le champ de mesure pour l'ensemble du corpus** 66
Encadré : à propos des documents exécutables 66
Récupération de corpus 66
Application des modèles à un nouveau corpus 67
Projection de la répartition 67
Préparation des données 67
Calcul de la marge d'erreur et des estimations 70
Extrapolation à l'ensemble du corpus. 72

**Phase 5 : dénombrement des images du portail Persée** 74
Objectifs 74
Construction de l'échantillon 74
Échantillonnage des images caviardées des collections rétrospectives de Persée 75
Échantillonnage des images des publications 2019 de Persée 79
Méthode d'analyse 79
Résultats de l'analyse 80
Résultat général 80
Résultats complémentaires 85
Dénombrement par extrapolation 87

# Contexte et structure de l'étude

**Les auteurs de publications scientifiques sont limités par les difficultés d'identification et d'obtention des droits sur les œuvres sur lesquelles leurs travaux s'appuient**, et par les coûts de transaction générés — à la fois le temps de recherche des ayants droits et le paiement de droits à la charge des éditeurs ou des auteurs —, excessifs au regard de l'économie de la publication scientifique dans les disciplines concernées, à savoir principalement en sciences humaines et sociales (SHS)[1]. En effet, l'*Étude sur l'économie des revues françaises en sciences humaines et sociales* réalisée par le Ministère de la culture en janvier 2020[2] montre qu'en moyenne, un numéro de revue est imprimé à 450 exemplaires. Ces difficultés font actuellement barrage à la diffusion sous forme numérique des travaux scientifiques qui s'appuient sur des corpus importants d'images, ce qui empêche leur mise à disposition en accès ouvert et réduit leur visibilité et leur impact.

Dans ce contexte, **une évolution du droit de la propriété intellectuelle vise à faciliter la reproduction d'œuvres protégées par le droit d'auteur dans les publications scientifiques** : l'article 28 de la loi de programmation de la recherche autorise le Gouvernement à prévoir, par voie d'ordonnance, l'octroi de licences collectives étendues permettant "l'utilisation d'œuvres relevant des arts visuels, à des fins exclusives d'illustration de publications, ou de travaux, diffusés en ligne sans restriction d'accès, dans le cadre d'une activité de recherche et d'enseignement supérieur publics, à l'exclusion de toute activité à but lucratif".

**Afin de préparer la mise en place des licences collectives étendues,** le Ministère de l'enseignement supérieur, de la recherche et de l'innovation a souhaité mener la présente étude visant à :

- établir, sur la base d'une connaissance de la volumétrie globale d'images publiées dans des travaux et publications scientifiques sur une année, une estimation du nombre d'images entrant dans le champ de la mesure ;
- fournir une méthodologie permettant de réitérer ce calcul sur les flux annuels de travaux et publications scientifiques, dans 3 ans, 5 ans, voire 10 ans si c'est encore possible ;
- évaluer le nombre d'images entrant dans le champ de la mesure dans les collections rétrospectives de publications scientifiques.

[1] Voir le rapport *Droits des images, histoire de l'art et société – Rapport sur les régimes de diffusion des images patrimoniales et leur impact sur la recherche, l'enseignement et la mise en valeur des collections publiques* publié en 2018 par l'Institut national d'histoire de l'art, disponible à l'adresse https://www.inha.fr/_attachments/de-nouvelles-democraties-du-savoir-actualite/rapport_images_usages221018.pdf

[2] Rapport final disponible à l'adresse https://www.culture.gouv.fr/content/download/262108/2986366

**L'étude est structurée comme suit** : la phase 1 vise à délimiter le corpus et dénombrer les images dans ce corpus. Les deux phases suivantes visent à caractériser finement les images qui relèvent du champ de la mesure, à partir d'un échantillon annoté manuellement (phase 2) puis en testant des méthodes automatiques qui doivent reproduire les résultats de l'analyse manuelle (phase 3). De retour sur le corpus, la phase 4 doit permettre de généraliser les résultats de l'échantillon et de proposer une méthode d'itération permettant chaque année de répéter l'analyse. La phase 5 complète le dénombrement des images avec la collection rétrospective de Persée.

L'ensemble de ces travaux ont été exécutés sous la direction du Comité de pilotage composé des membres suivantes :

**Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation :**

- Odile Contat, département de l'information scientifique et technique et du réseau documentaire
- Marin Dacos, conseiller pour la science ouverte auprès du directeur général de la recherche et de l'innovation
- Francis Prost, chargé de mission archéologie, mondes anciens au service de la stratégie de la recherche et de l'innovation (secteur sciences de l'homme et de la société)
- Claire Leymonerie, chargée de mission pour la science ouverte à la direction de la recherche et de l'innovation

**Ministère de la Culture** :

- Hugues Ghenassia de Ferran, secrétariat général, sous-direction des affaires juridiques
- David Pouchard, secrétariat général, sous-direction des affaires juridiques
- Chantal Devillers-Sigaud, Direction générale de la création artistique (DGCA), chargée de mission au bureau des affaires juridiques
- Ludovic Julié, DGCA, chargé de mission auprès du délégué aux arts visuels
- Marion Hislen, DGCA, cheffe du bureau de la photographie
- Julie Franc, Direction générale des médias et industries culturelles, sous-direction de la presse et des métiers de l'information
- Anne Dubile, Direction générale des patrimoines et de l'architecture, Service des Musées de France, cheffe adjointe du bureau du pilotage des musées nationaux

**Organismes de gestion collective** :

- Thierry Maillard, directeur juridique de l'ADAGP
- Serge Monnet, responsable informatique à l'ADAGP
- Agnès Defaux, directrice juridique de la SAIF

**Experts** :

- Emmanuelle Bermès, adjointe scientifique et technique au directeur des services et des

réseaux de la BnF

- Patrick Peccatte, chercheur associé au laboratoire d'histoire visuelle contemporaine

# Phase 1 : délimitation du corpus et dénombrement des images

## Extraction des documents

### La récupération des métadonnées sur Isidore

Conformément au cahier des charges, **la phase 1 s'appuie sur la base Isidore pour constituer un corpus des publications françaises en accès ouvert sur une année**. Isidore est un moteur de recherche spécialisé dans les sciences humaines et sociales, développé par la très grande infrastructure de recherche Huma-num : https://isidore.science/ Créé en 2009, Isidore indexe aujourd'hui près de 9 millions de documents issus de différentes plateformes de publication scientifiques ou d'archive ouverte. Couvrant originellement la production française, Isidore s'est progressivement ouvert à l'international.

La démarche générale consiste à remonter de l'aval (les publications comportant des images identifiées via Isidore), vers l'amont, en récupérant les publications identifiées, puis en extrayant les images de ces publications. De cette manière sera constitué le corpus d'images qui servira de base à une estimation de la volumétrie globale d'images concernées par la mesure et à l'échantillonnage en vue de l'analyse documentaire de la phase 2.

Les principales plateformes académiques indexées par Isidore en termes de volume sont :

- OpenEdition, plateforme qui concentre quatre espaces de publication en accès ouvert : revues scientifiques (OpenEdition Journals), carnets de recherche (Hypothèses), livres (OpenEdition Books) et annonces (Calenda);
- les plateformes d'archives ouvertes du Centre pour la communication scientifique directe (CCSD), dans lesquelles les chercheurs ou leurs institutions déposent des articles (en version auteur), des thèses ou d'autres documents. Il s'agit en particulier de HAL (archive ouverte pluridisciplinaire de la recherche française), de TEL (thèses de doctorat et habilitations à diriger des recherche) et de MediHAL (archive ouverte qui permet de déposer des données visuelles et sonores produites dans le cadre de la recherche scientifique);
- Thèses.fr, portail de référencement des thèses de l'Agence bibliographique de l'enseignement supérieur (ABES);
- Cairn, plateforme de publication et de livres en sciences humaines et sociales, essentiellement en accès payant ou bien en accès ouvert après application d'une barrière mobile (par exemple, si la barrière mobile est de 3 ans, seules les publications parues à l'année n-3 seront en accès ouvert, les publications plus récentes étant en accès payant).
- Persée, portail donnant accès à des collections rétrospectives de revues et livres scientifiques numérisés, essentiellement en sciences humaines et sociales.

L'extraction a été réalisée à partir du « point SPARQL ». Il s'agit d'une infrastructure de requête

s'appuyant sur le web sémantique qui permet d'exploiter efficacement le schéma de base de données d'Isidore. Isidore est en effet un moteur de recherche très documenté qui ne donne pas seulement accès aux métadonnées fondamentales (titre, auteur, date…) mais aussi à des informations documentaires explicitement pensées pour un usage de recherche en SHS (discipline, type du document, nom de la revue…) Plusieurs extractions successives ont pu être réalisées à partir de requêtes différenciées :

- l'ensemble des documents d'Isidore pour l'année 2019 et pour l'année 2017 (pour le corpus Cairn).
- les types et les disciplines des documents d'Isidore pour l'année 2019.
- les métadonnées des « collections » d'Isidore qui renvoient en réalité à n'importe quel ensemble documentaire cohérent : une revue, un carnet de recherche ou une base de données ou de notices bibliographique.

Cependant, Isidore n'indique pas deux informations importantes pour l'étude : la nationalité de l'éditeur et la disponibilité du document en libre accès.

## Délimitation du corpus de documents

**L'étude s'appuie sur un corpus hybride** comprenant initialement :

- l'ensemble des publications de 2019 hors Cairn (soit 266 069[3] références)
- les publications Cairn de 2017 (soit 17 109 références) plutôt que 2019 afin d'intégrer les publications ouvertes après expiration de la barrière mobile (en effet, les publications datées de 2019 sur Cairn sont majoritairement payantes et donc hors du champ de la mesure ; or il est important d'analyser également les pratiques de publication des revues disponibles sur Cairn, d'où le choix d'une année de publication plus ancienne car après 3 ans ces articles passent typiquement en accès ouvert).

Dans la phase 5, nous ajouterons l'intégralité des publications disponibles dans Persée.

Qu'appelle-t-on une publication dans ce décompte ? Selon la classification d'Isidore, le corpus comprend en majorité :

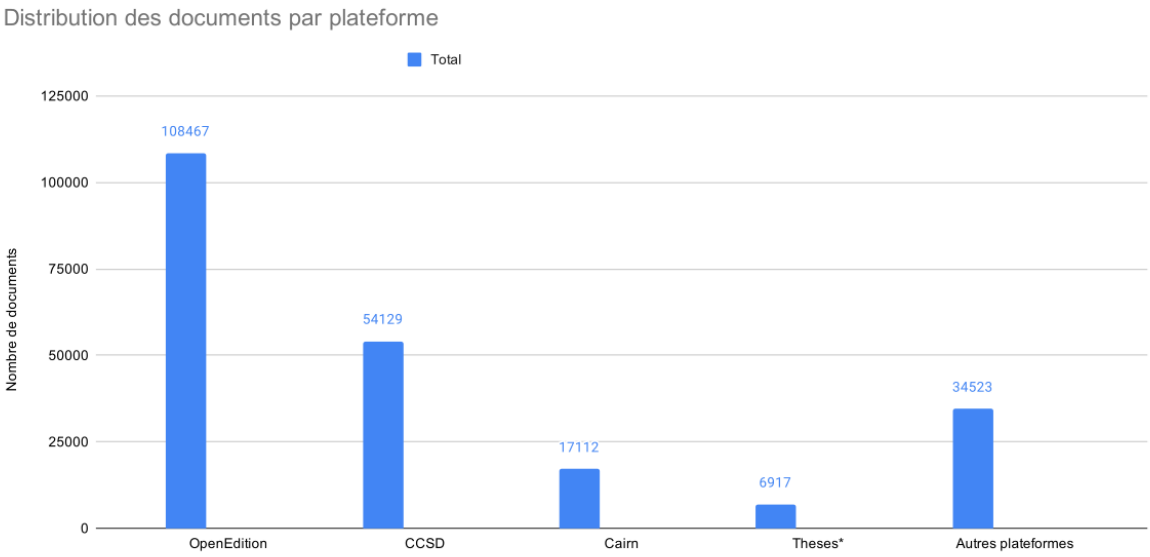
- des articles de revues (OpenEdition Journals, Cairn) qui peuvent aussi avoir été déposés dans des archives ouvertes (HAL)
- des chapitres d'ouvrages (OpenEdition Books) ; en publication numérique le chapitre devient l'unité de référence plutôt que le livre
- des thèses et mémoires d'habilitation à diriger des recherches (Theses.fr, HAL)
- des billets de blog (Hypothèses)

En accord avec le comité de pilotage, les résultats d'Isidore n'ont pas été dédoublonnés, même si une publication peut être disponible sur plusieurs plateformes (parfois avec des différences). C'est un artefact des pratiques de publication, et le dédoublonnage serait trop coûteux pour

[3] Le chiffre précédemment mentionné de 277 500 références incluait des doublons d'Isidore issus de corpus étrangers sans incidence sur les corpus de collections françaises.

pouvoir être itéré facilement.

À cette étape le corpus comprend donc 266 069 + 17 109 soit 283 178 références[4]. En pratique, **quatre grandes plateformes de publications scientifiques concentrent la plupart des publications du corpus** : OpenEdition, le CCSD (notamment HAL…), Cairn et Theses.fr. Voici les volumes correspondant à ces quatre plateformes :




**Sur ce corpus de 283 178 références, nous appliquons un premier filtre relatif à la nationalité**. En effet, **le Comité de pilotage convient que la mesure ne concerne que les textes édités en France. Pour ce faire, nous retenons la nationalité de l'éditeur** et non de l'hébergeur :

- pour les plateformes hébergeant des revues et des ouvrages, nous avons identifié la nationalité de l'éditeur (et non de l'hébergeur) à partir des données de référence d'ISSN
- pour les carnets de recherche sur Hypothèses.org, nous avons eu accès aux données internes d'OpenEdition, sachant que chaque carnet fait l'objet d'une demande de création, laquelle comprend le pays de rattachement du carnet
- sur les autres plateformes, la nationalité des documents a été considérée en bloc ; par exemple les documents déposés sur HAL ont été considérés comme français, les thèses recensées sur theses.fr ont été considérées comme françaises, alors que les archives ouvertes des universités de Liège et de Genève, également indexées dans Isidore, n'ont pas été considérées comme françaises.

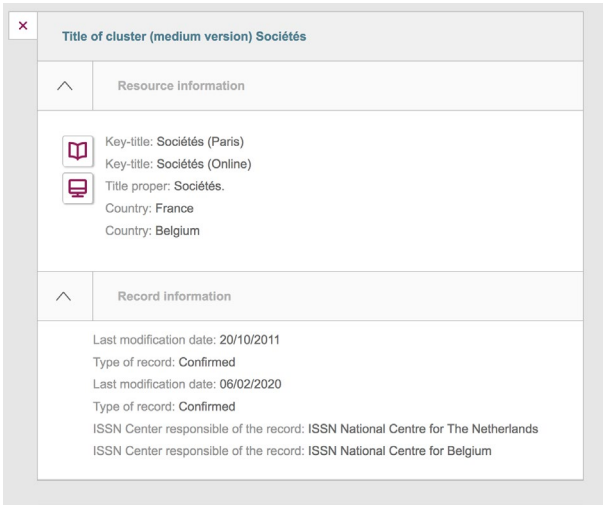
À titre d'illustration de ce filtre appliqué à l'échelle de la revue et non de la plateforme, nous excluons plusieurs dizaines de revues belges de la plateforme française Cairn mais incluons 5 revues françaises du portail de publication scientifique canadien Érudit, à savoir :

- *Revue internationale P.M.E.: Économie et gestion de la petite et moyenne entreprise*
- *Études littéraires africaines*
- *Bulletin de la Société d'Histoire de la Guadeloupe*

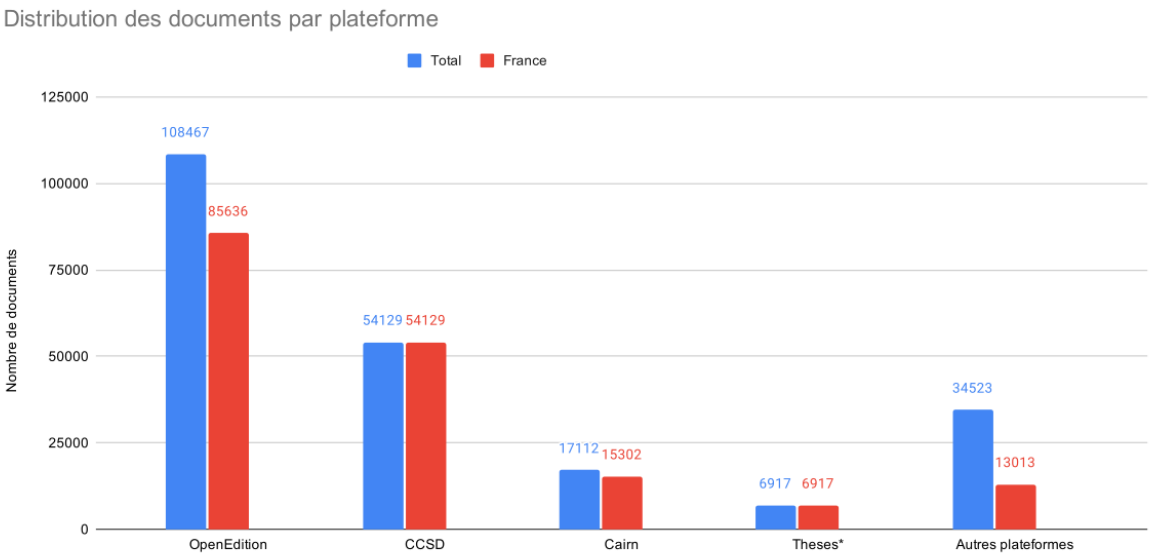
[4] L'ensemble des métadonnées sont consultables à l'adresse : https://images-publications-scientifiques.huma-num.fr/livrable/isidore_meta.zip

- *Revue des sciences de l'eau / Journal of Water Science*
- *Phronesis*

**Dans certains cas limite, la détermination de la nationalité est relativement incertaine** : sur Cairn, l'éditeur belge De Boeck a acquis plusieurs revues françaises et les ISSN imprimés se trouvent attribués en France tandis que les ISSN numériques sont attribués en Belgique (voir capture ci-dessous). Avec l'accord du Comité de pilotage nous avons considéré comme françaises ce type de publications bi-nationales.

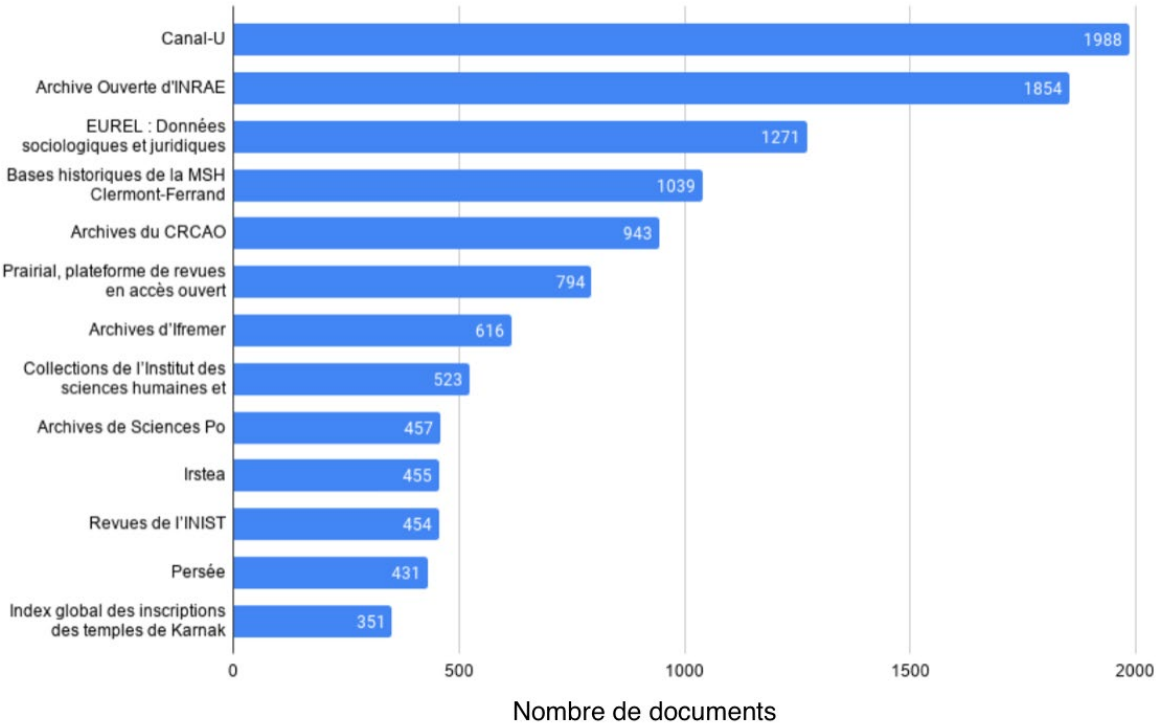



**Après application du filtre de nationalité, la taille du corpus est réduite à 176 787 références au total**. Voici comment le corpus évolue pour les quatre plateformes majeures, par comparaison avec la première étape (sans le filtre, en bleu) :




Les autres plateformes et sites indexés dans Isidore représentent collectivement un peu moins de 10% du corpus (13 013 documents sur 176 787). Voici le haut de la distribution de ces autres plateformes, qui suit une loi de puissance typique en bibliométrie (loi de Bradford) et traduit

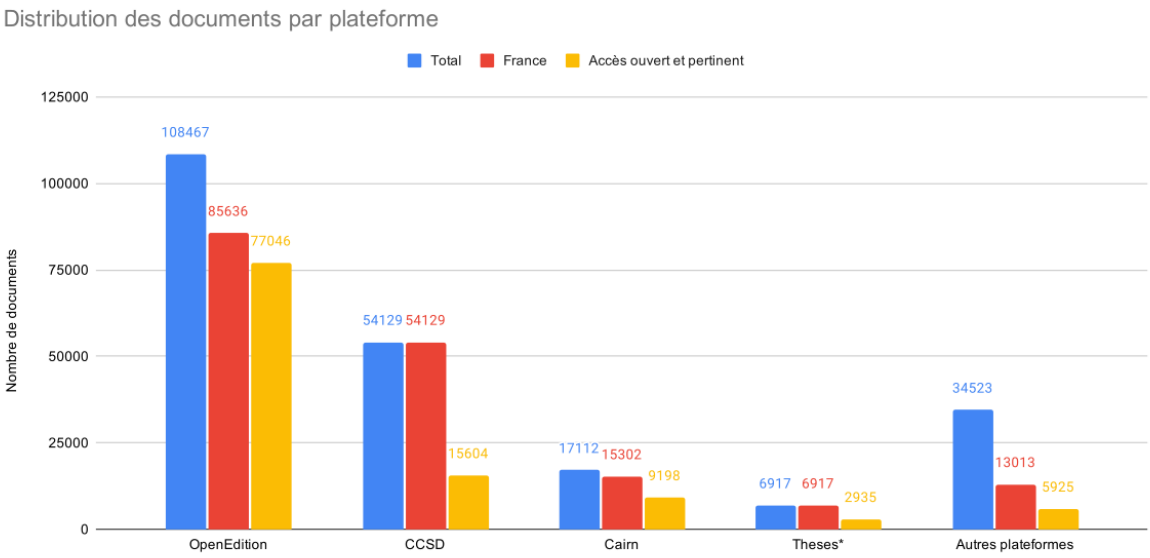
la concentration des documents entre les sources :




Ces plateformes sont bien prises en compte dans l'étude, **à l'exception de Canal-U dont le Comité de pilotage convient qu'elle n'est pas pertinente puisqu'elle héberge des vidéos**. Nous appliquons donc un deuxième filtre qui coïncide avec la catégorie "Image en mouvement" d'Isidore, soit 1 988 références (voir graphique ci-dessus). Par ailleurs plusieurs collections n'étaient en réalité que des miroirs de HAL (les Archives du CRCAO, l'Institut des sciences humaines et sociales du CNRS)

Enfin, étant donné que la mesure ne concerne que les documents en accès ouvert (accès gratuit, sans restriction d'accès technique), le troisième et dernier filtre appliqué permet **d'exclure les références dont le texte intégral ne serait pas disponible, ou seulement après paiement. Le corpus final comprend actuellement 110 708 documents**. Le critère "accès ouvert" n'est pas un critère binaire fourni par Isidore et a dû être évalué au cas par cas.

Parmi les quatre plateformes principales, ce sont les plateformes du CCSD qui ont la plus faible proportion de documents en accès ouvert (c'est-à-dire une majorité de notices bibliographiques sans contenu associé) comme l'indique le graphique ci-dessous :




À noter qu'il y a un recoupement de 795 documents entre le CCSD et theses.fr ce qui porte le total réel des thèses à 3 730 documents. Pour ne pas fausser les chiffres généraux, nous n'avons pas téléchargé de nouveau ces documents déjà présents dans le corpus du CCSD.

Les autres collections subsidiaires comprennent 5 438 documents (dont 4 068 au format HTML et 1 857 documents au format PDF) auxquels s'ajoutent 487 de Persée qui restent encore à obtenir et seront analysés dans le cadre de la phase 5 de l'étude. Ces résultats confirment l'hypothèse initiale d'une forte concentration de la publication SHS en accès ouvert et justifient de traiter en priorité les quatre grandes plateformes.

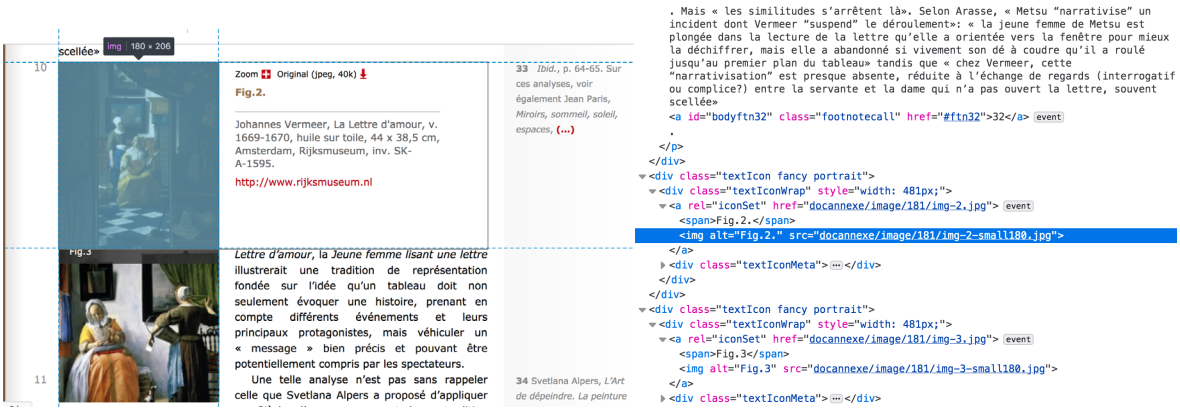
## Extraction des images

Le corpus est principalement disponible sous deux formes : des documents web en langage HTML et des fichiers PDF proches du format de l'imprimé. À titre d'exemple, les revues en ligne (OpenEdition, Cairn…) privilégient le HTML tandis que les archives ouvertes hébergent des fichiers PDF. Le processus d'extraction des images sera complètement différent dans les deux cas.

**L'extraction a produit un corpus préliminaire de 915 296 *objets visuels* dont 899 567 objets « valides »[5] qui ne sont pas tous des images.** Comme nous le détaillons dans chaque section, le processus d'extraction a retourné un nombre important d'artefacts et de fichiers numériques qui ne sont en réalité pas des illustrations. Les outils de classification automatiques décrits dans la section suivante seront déterminants pour permettre d'établir une estimation générale du nombre d'images en SHS pendant l'année 2019.

### Corpus HTML

Nous avons récupéré l'ensemble des balises "images" ou `<img>` dans les documents nativement web texte principal des documents. Il est nécessaire de bien cibler le corps du texte pour ne pas récupérer des éléments génériques du site (logo par exemple). Si l'inclusion d'une image (ou d'autres éléments détournés comme des images) correspond à un standard universel, ce n'est pas le cas des articles et des productions scientifiques. D'un site à l'autre les conventions varient plus ou moins significativement. L'article peut être inclus dans une balise `<article>` mais aussi dans une balise « divers » `<div>` avec des identifiants personnalisés (id, main, content…). Voici un exemple d'image identifiée sur un article OpenEdition et sa récupération via le code source de la page :




[5] Un petit nombre d'images ne sont pas utilisables, soit parce que trop dégradées, soit parce que tout simplement vides (message d'erreur « Image data missing »). Elles représentent environ 1,5% du corpus visuel initial.

En pratique, nous avons pu constater que la balise `<img>` est fréquemment détournée pour contourner des limitations diverses liées à la publication en ligne. Les tableaux inclus dans les articles de recherche d'OpenEdition Journals et d'OpenEdition Books apparaissent ainsi fréquemment comme des images, bien qu'il existe des balises dédiées pour ce format en HTML (`<table>` et ses subdivisions). La création d'un tableau HTML nécessite un travail d'adaptation plus conséquent que la capture d'écran d'un tableau créé sur Word. De plus, certains symboles et certaines expressions ne peuvent pas être aisément générées avec les outils d'éditions disponibles : c'est particulièrement le cas des équations et des notations mathématiques.

D'après les résultats de la classification automatique **près d'un tiers des images mises en ligne ne sont en réalité pas des « images » au sens usuel du terme** mais des détournements de la balise `<img>`.

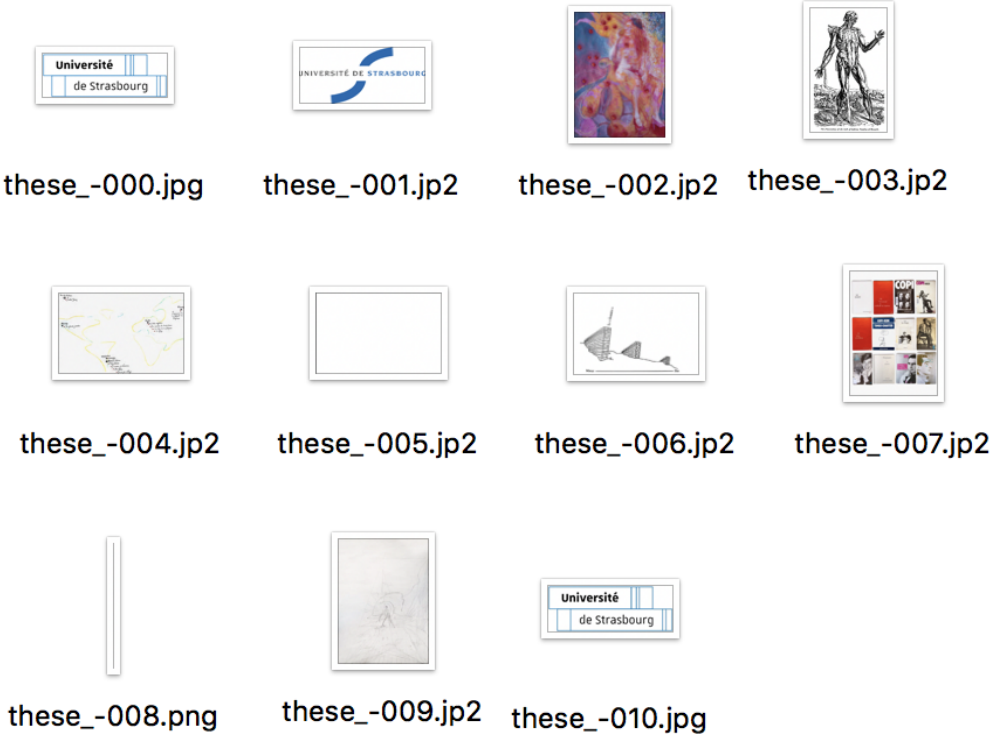
## Corpus PDF

L'extraction des images de documents PDF est plus complexe. Il n'y a pas de balises images mais des documents images directement inclus à l'intérieur du document. Pour les identifier et les extraire nous avons utilisé le programme `pdfimages`, qui permet d'obtenir le résultat ci-dessous. Chaque ligne du tableau correspond à une image et à ses métadonnées qui incluent notamment sa résolution et sa position dans le document (page, coordonnées…).

```
MBPdePierreCarl:Word2Entity pierre-carllanglais$ pdfimages -list these_internet_zouari_n.pdf
page   num  type   width height color comp bpc  enc interp  object ID x-ppi y-ppi size ratio
--------------------------------------------------------------------------------------------
   1     0 image     975   547  icc     3   8  jpeg   no      1571  0   583   583  121K 7.7%
   1     1 image     229   265  icc     3   8  image  no      1572  0  2069  2069 4238B 2.3%
   1     2 smask     229   265  gray    1   8  image  no      1572  0  2069  2069  261B 0.4%
   2     3 image     437   215  icc     3   8  jpeg   no      1577  0   223   224 13.3K 4.8%
   2     4 image     463   261  icc     3   8  jpeg   no      1576  0   300   300 9.82K 2.8%
  22     5 image    1594   580  icc     3   8  jpeg   no      1641  0   367   367 39.1K 1.4%
  22     6 image    1594   580  icc     3   8  jpeg   no      1640  0   367   367 69.0K 2.5%
  22     7 image    1594   580  icc     3   8  jpeg   no      1639  0   367   367 47.5K 1.8%
  22     8 image    1594   578  icc     3   8  jpeg   no      1638  0   367   366 41.4K 1.5%
  86     9 image     686   685  icc     3   8  jpeg   no      1834  0   175   175 39.0K 2.8%
  88    10 image     921   654  icc     3   8  jpeg   no      1841  0   234   234 47.4K 2.7%
  91    11 image    3922   259  index   1   8  image  no      1857  0   665   665  454K  46%
  91    12 image    3922   259  icc     3   8  jpeg   no      1856  0   665   665  108K 3.6%
  91    13 image    3922   259  icc     3   8  jpeg   no      1855  0   665   665  119K 4.0%
  91    14 image    3922   259  icc     3   8  jpeg   no      1854  0   665   665  130K 4.4%
  91    15 image    3922   259  icc     3   8  jpeg   no      1853  0   665   665  108K 3.6%
  91    16 image    3922   259  icc     3   8  jpeg   no      1868  0   665   665 91.1K 3.1%
  91    17 image    3922   259  icc     3   8  jpeg   no      1866  0   665   665  104K 3.5%
  91    18 image    3922   259  icc     3   8  jpeg   no      1864  0   665   665  111K 3.7%
  91    19 image    3922   251  index   1   8  image  no      1862  0   665   661  642K  67%
  91    20 image    4160   234  index   1   8  image  no      1861  0   705   706  584K  61%
  91    21 image    4160   234  icc     3   8  jpeg   no      1860  0   705   706  137K 4.8%
  91    22 image    4160   234  icc     3   8  jpeg   no      1859  0   705   706  132K 4.6%
  91    23 image    4160   234  icc     3   8  jpeg   no      1858  0   705   706  114K 4.0%
```

Quantitativement, l'extraction a concerné principalement les collections du CCSD (15 604 documents), theses.fr (2 935 documents) et quatre collections secondaires (Ifremer, les revues de l'Inist, la Bibliothèque numérique du Service Régional de l'Archéologie de Bretagne et Toulouse Capitole Publications).

Les résultats sont imparfaits. Non seulement l'extraction inclut des images hors textes (logos en page de garde) mais aussi de nombreux artefacts : des tableaux, des morceaux de texte et des objets « vectoriels ». Les dessins créés avec certains logiciels d'image (comme Inkscape) ne sont pas considérés comme des images mais comme des agrégats d'éléments dessinés (lignes, formes, lettres…). Selon leur processus d'édition, certains documents apparaissent en réalité comme une collection d'images de textes.




Lors du traitement des 15 604 documents du CCSD, nous avons d'abord obtenu près de 1,5 millions d'images : la grande majorité de ces images correspondent à des artefacts, certains documents comprenant plusieurs dizaines de milliers d'images consistant en une série de fragments de textes et d'éléments visuels microscopiques.

Nous avons pris la décision de retirer toutes les images d'une taille inférieure à 1,5 kilo-octets. Cette approche permet de diminuer une grande partie du bruit sans affecter aucune image pertinente pour l'étude : avec une taille de 1,5 ko, la résolution est beaucoup trop faible pour une vraie illustration, en dehors de petits logos ou pictogrammes. Dans le corpus d'images utilisé pour créer le modèle de classification décrit dans la section suivante, aucune image n'a une taille inférieure à 2 ko.

Dans le cas du CCSD, le nombre d'images est passé de 1,5 millions à 375 000. Une grande partie de ce corpus visuel reste composé d'artefacts. Si les éléments vides ou quasi-vides ont disparu, il reste ainsi de nombreux fragments de textes suffisamment "riches" pour être retenus par l'extraction.

Le Comité de pilotage peut déterminer un seuil plus élevé s'il estime qu'une image trop petite n'est pas suffisamment reconnaissable au sens de la loi. Nous avons extrait à cette fin des

échantillons :

- d'images comprises entre 1,5 ko et 2 ko : https://docs.google.com/spreadsheets/d/1ktYCFnXVbFMMvVeGll2PrBHDOUYrN7HC-15P6qRNS1Y/edit?usp=sharing
- d'images comprises entre 2 ko et 10 ko : https://docs.google.com/spreadsheets/d/1JeA8HxhEYE3NZN7UzpNcA4mMHIHmmjNETKnqDCwyZr4/edit?usp=sharing
- d'images comprises entre 10 ko et 30 ko : https://docs.google.com/spreadsheets/d/11y-I6B3ob6vNi9fNg-pm5nxE_c0ZZDRxuJjncnr3ujc/edit?usp=sharing

Pour chaque échantillon, l'image concernée apparaît dans la colonne D et peut être visionnée dans son contexte en cliquant sur le lien vers le document dans la colonne C.

## Classification des images

**Le corpus initial de 899 567 objets visuels a été classé automatiquement avec un modèle de *deep learning*.** La classification ne permet pas seulement de décrire automatiquement les images mais aussi de constituer un corpus d'images pertinent pour l'étude. Comme nous avons pu le voir dans les sections précédentes, le format « image » est fréquemment détourné pour représenter des objets qui ne sont pas des illustrations (logos, tableaux, équations) et l'extraction des images contenues dans les PDF reste imparfaite et inclut de nombreux artefacts (éléments vectoriels, textes…). Pour toutes ces raisons, nous proposons d'utiliser la classification pour estimer le nombre d'images au sens usuel du terme.

La classification repose sur un modèle *resnet 50* ré-entraîné sur un corpus de 4 890 images issues des quatres grandes plateformes de notre corpus (OpenEdition, archives ouvertes du CCSD, Cairn, theses.fr) et d'un corpus externe WikiPainting. Cette méthode de *transfer learning* permet de créer des modèles de classification à moindre coût : au lieu de créer un modèle inédit, ce qui est très coûteux sur le plan computationnel, nous partons d'un modèle existant capable d'identifier environ 1 000 objets. En quelque sorte, le modèle dispose déjà une culture visuelle généraliste qui lui permet d'assimiler plus facilement les formes et les représentations propres aux publications francophones en sciences humaines et sociales.

Le modèle fait partie des livrables de l'étude[6].

### Catégories du modèle

[6] Les modèles "model_imsci.pt" et "model_imsci_pdf.pt" sont disponibles sur un dépôt de codes sources gitlab : https://gitlab.huma-num.fr/planglais/images-usages-isidore/-/tree/main/modeles La version PDF est quasiment identique et intègre des exemples supplémentaires d'artefacts qui ne se trouvent que dans les corpus de documents PDF.

Le modèle comprend 35 catégories. Toutes les catégories ne sont pas utiles pour l'étude mais contribuent à améliorer la qualité et l'efficacité du modèle. Par exemple, il aurait été possible de faire une catégorie fourre-tout comprenant tous les éléments d'emblée non pertinents comme les textes, les tableaux, les logos ou les équations ; mais moins une catégorie est associée à une représentation précise et moins elle est opérationnelle, avec le risque de retenir davantage de faux positifs ou inversement de manquer davantage d'images qui relèvent de la classification. La création de catégories fines, parfaitement adaptées à la morphologie du corpus, rend également le modèle réutilisable pour d'autres études ultérieures.

Les catégories ont été élaborées empiriquement à partir de la consultation du corpus d'entraînement aléatoire.

Elles comprennent d'abord des **catégories non pertinentes** que nous cherchons à exclure du corpus ou qui feront l'objet d'un traitement à part :

- les tableaux de données
- les éléments textuels : mots, phrases, paragraphes…
- les lettres isolées ; il s'agit principalement d'artefacts ou de symboles qui ne peuvent pas être aisément figurés dans le corps du texte
- des artefacts divers issus de l'extraction des images dans les PDF
- les images vides
- les équations
- les logos, très présents sur les pages de garde des documents du CCSD et des thèses avec logo de la plateforme (HAL, theses.fr, Dumas…) et/ou logos d'universités et organismes de tutelle
- les mentions normalisées d'image caviardées ; dans l'attente du corpus Persée, cela ne comprend que la mention « Droits numériques non obtenus » utilisée sur les plateformes d'OpenEdition.

Plusieurs catégories portent sur des **visualisations de données dont on suppose qu'elles sont produites par les auteurs du texte.** Le travail de documentation de l'échantillon en phase 2 permettra de vérifier si c'est bien systématiquement le cas :

- les graphiques : graphes linéaires, camemberts, histogrammes en bâton
- les coupes géologiques représentant les strates successives d'un terrain, une représentation courante en géographie
- les plans de maisons et d'habitations souvent représentés en archéologie
- les schémas, qui comprennent typiquement des éléments textuels associés par des flèches
- les cartes géographiques
- les graphes de réseaux, utilisés pour représenter des liens entre de nombreuses entités représentées en deux dimensions.

Nous avons aussi inclus des catégories décrivant **des reproductions de documents** :

- les couvertures de livres et d'ouvrages (fréquemment reproduites dans les recensions publiées dans les carnets de recherche et les revues)
- les posters et affiches, ce qui inclut notamment les affiches de journées d'étude et de colloque
- les inscriptions anciennes (principalement en grec et en latin)
- les reproductions de textes manuscrits
- les reproductions de pages entières de documents divers (livres, journaux…)
- les reproductions de pages entières de documents scientifiques ; cette catégorie vise surtout à identifier une erreur récurrente de l'extraction des images dans les PDF : la page entière apparaît comme une image (et ne constitue donc pas un objet pertinent)
- les reproductions de pièces de monnaie, assez fréquentes dans les revues d'histoire antique et médiévale
- les captures d'écrans de logiciels ou de sites en ligne.

Il convient de noter que l'on retrouve dans cette catégorie les panels d'images, où plusieurs illustrations sont assemblées par les auteurs au sein d'une même image. À ce titre, le dénombrement des images dans cette catégorie n'est pas sans ambiguïté.

Enfin, le dernier ensemble de catégories recouvre les **images figuratives, qui constituent un champ d'investigation privilégié de l'étude** :

- les photographies de paysages urbains et naturels
- les photographies d'espaces intérieurs
- les portraits photographiques
- les photographies de « scènes » (interactions entre plusieurs personnes)
- les photographies d'objets
- les dessins de tous types ; cette catégorie est assez hétérogène et comprend aussi bien les croquis d'artefacts (dans les travaux d'archéologues) que les reproductions de bandes dessinées
- les sculptures

Dans le corpus servant à entraîner le modèle, la part des œuvres relevant des arts visuels est faible, de l'ordre de moins de 5% de l'ensemble. Pourtant il s'agit d'une catégorie importante pour l'étude et plus particulièrement pour la phase ultérieure d'annotation manuelle. Pour cette raison, nous avons décidé de **compléter le corpus d'entraînement avec un corpus complémentaire d'œuvres visuelles externe issu de la base WikiPainting**[7]. Cette base comprenant près de 100 000 peintures extraites de Wikimédia Commons est communément utilisée pour créer des modèles de deep learning en art visuel. Chaque œuvre est documentée

[7] Le corpus et les métadonnées associées sont disponibles sur Kaggle : https://www.kaggle.com/c/painter-by-numbers/code

selon deux catégories : le genre (portrait, peinture historique, allégorie, peinture abstraite) et le mouvement (néo-classicisme, impressionnisme, renaissance du Nord, etc.).

Nous avons créé une version simplifiée du corpus de *WikiPainting* en ne retenant que cinq catégories :

- les portraits
- les paysages
- les « scènes » (qui agrègent les peintures de genre),
- les représentations d'objets inanimés et natures mortes
- la peinture abstraite.

Chacune de ces sous-catégories a été enrichie avec les illustrations présentes dans le corpus d'images scientifiques.

La composition du corpus d'entraînement donne un bon aperçu des usages visuels sur les plateformes d'édition scientifique en HTML comme Cairn ou OpenEdition. Les images ont en effet été extraites de manière aléatoire à partir de l'ensemble du corpus. Elles se répartissent en quatre grandes catégories : les éléments qui ne sont pas des images (29,31% : textes, tableaux, lettres, artefacts, logos…), les cartes, graphiques et schémas (31,5%), les reproductions de documents (14,2% : couvertures de livres, sites web, manuscrits…) et les représentations figurées (25,05% : peintures, photographies, dessins). Seule la dernière catégorie relève sans ambiguïté de la mesure.

## Évaluation du modèle

Le modèle a été entraîné avec l'application Google Colab. Il n'a pas été possible d'utiliser R Studio pour cette tâche en raison des limitations de l'environnement fourni par Huma-Num (absence d'un processeur graphique ou GPU).

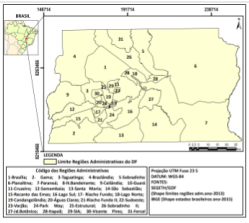
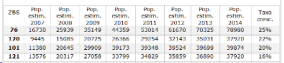
Nous avons testé manuellement les résultats du modèle sur un corpus aléatoire de 1 000 images d'OpenEdition. Le taux d'erreur est de l'ordre de 5% entre les grandes catégories (éléments non pertinents, visualisation de données, documents, images figuratives), c'est-à-dire que 5 images sur 100 ne sont pas classées correctement.

Le modèle continuera d'être affiné lors des deux phases suivantes de l'étude et bénéficiera notamment des retours de l'annotation de l'échantillon de la phase 2.

## Résultats de la classification

La base de métadonnées fournie avec les livrables donne pour chaque image du corpus la catégorie la plus probable et la probabilité associée. L'extraction ci-dessous permet d'avoir une

idée générale de la qualité du modèle[8]. Le tableau peut se lire ainsi : la première image a 77% de chances d'être un document, la suivante a 70% de chances d'être une carte, et la troisième a 95% de chances d'être un tableau.

| Identifiant | Lien du docum | Lien Isidore | Image | Classification | Probabilité | I |
|---|---|---|---|---|---|---|
| 1.kvu5t3_progra | http://travailform | http://isidore.scie |  | page_document | 77.36270905 | I |
| confins.22922_ir | http://journals.op | http://isidore.scie |  | map | 70.91641235 | I |
| confins.22922_ir | http://journals.op | http://isidore.scie |  | table | 95.04145813 | I |

En filtrant par probabilités il est possible de se focaliser uniquement sur les corpus les plus pertinents. Les photographies de paysage, avec une probabilité supérieure à 90%, relèvent toutes de cette catégorie.

Le tableau suivant donne les résultats de la classification pour les quatre grandes plateformes ainsi que pour les autres ressources au format HTML et les ressources au format PDF. Le tableau peut se lire ainsi, en prenant l'exemple de la première cellule : le corpus OpenEdition comprend 16 071 images de type tableaux, textes… qui représentent 18% des images de cette collection. La somme des pourcentages en ligne vaut toujours 100% (aux erreurs d'arrondi près).

| Collection | Tableaux, textes, logos et artefacts | Graphe, cartes et schémas | Documents | Images figuratives |
|---|---|---|---|---|
| *OpenEdition* | 16 071 (18%) | 23 071 (25%) | 16 072 (17%) | 39 628 (40%) |
| *Theses* | 189 768 (55%) | 94 955 (28%) | 16 713 (5%) | 43 775 (13%) |
| *CCSD* | 186 520 (50%) | 86 005 (23%) | 29 998 (8%) | 74 014 (20%) |

[8] Document disponible à l'adresse https://docs.google.com/spreadsheets/d/1QdfoD6txGFG0bbynRgsKNGvias-l2blhBvr8GrBEO00/edit?usp=sharing

| | | | | |
|---|---|---|---|---|
| *Cairn* | 7 731 (51%) | 3 718 (25%) | 1 066 (7%) | 2 625 (17%) |
| *Autres corpus en PDF* | 28 140 (45%) | 15 627 (26%) | 4 438 (7%) | 13 088 (22%) |
| *Autres corpus en HTML* | 732 (25%) | 396 (14%) | 424 (15%) | 1 336 (46%) |
| *Total* | 428 962 (48%) | 223 772 (25%) | 68 711 (8%) | 174 466 (20%) |

Nous pouvons aussi « filtrer » graduellement le corpus pour s'approcher des catégories les plus pertinentes pour l'étude. Dans le tableau suivant, la troisième colonne « images » exclut les tableaux, textes, logos et autres artefacts. La dernière colonne concerne uniquement les images figuratives (photographies et peintures).

| **Collection** | **Objets visuels identifiés** | **Objets visuels valides** | **Images** | **Images figuratives** |
|---|---|---|---|---|
| *OpenEdition* | 96719 | 94850 | 78770 | 39628 (40%) |
| *Theses* | 356572 | 345211 | 162179 | 43775 (13%) |
| *CCSD* | 378573 | 376710 | 190017 | 74014 (20%) |
| *Cairn* | 15435 | 15140 | 7409 | 2625 (17%) |
| *Autres plateformes en PDF* | 65119 | 64788 | 33153 | 13088 (22%) |
| *Autres plateformes en HTML* | 2878 | 2868 | 2156 | 1336 (46%) |
| **Total général** | 915296 | **899567** | **466949** | **174466 (20%)** |

Nous donnons ci-dessous une version inversée du tableau précédent afin de mieux évaluer la place respective des différentes plateformes dans les différentes catégories d'image.

| **Collection** | *OpenEdition* | *Theses* | *CCSD* | *Cairn* | *Autres plateformes en PDF* | *Autres plateformes en HTML* |
|---|---|---|---|---|---|---|
| **Objets visuels identifiés** | 96719 | 356572 | 378573 | 15435 | 65119 | 2878 |
| **Objets visuels valides** | 94850 | 345211 | 376710 | 15140 | 64788 | 2868 |
| **Images** | 78770 | 162179 | 190017 | 7409 | 33153 | 2156 |
| **Images figuratives** | 39628 (23%) | 43775 (25%) | 74014 (42%) | 2625 (2%) | 13088 (8%) | 1336 (1%) |

Ces chiffres doivent évidemment être pris comme des estimations. **Avec un taux d'erreur de 5%, le nombre d'images figuratives d'OpenEdition se situe dans une fourchette 37 500 – 41 500.**

Ces résultats pourront être affinés par des perfectionnement ultérieurs du modèle, mais les ordres de grandeur resteront les mêmes.

La classification met en évidence que la part prépondérante des images issues du CCSD et de theses.fr résulte en grande partie d'artefacts. Nous trouvons 43% d'images figuratives dans OpenEdition mais seulement 20% pour le CCSD et 13% pour theses.fr. En ne prenant que les images figuratives, ces trois grandes plateformes occupent une place comparable dans le corpus. Cairn occupe une position plus secondaire : comme nous avons pu le voir, une grande partie des illustrations sont en réalité composées de texte (tableaux, équations…).

Le troisième tableau présente la distribution des classifications d'images par type de document (selon la nomenclature fournie par Isidore) :

| Type | Tableaux, textes, logos et artefacts | Graphiques, cartes et schémas | Documents | Images figuratives |
|---|---|---|---|---|
| *Article de revue* | 30556 (56%) | 13521 (25%) | 3319 (6%) | 6751 (12%) |
| *Blog* | 3992 (12%) | 2593 (8%) | 10364 (32%) | 15709 (48%) |
| *Chapitre* | 16807 (32%) | 15201 (29%) | 4112 (8%) | 16462 (31%) |
| *Livre* | 8004 (49%) | 2450 (15%) | 1366 (8%) | 4610 (28%) |
| *Mémoire de master* | 18358 (34%) | 8915 (17%) | 10524 (20%) | 15778 (29%) |
| *Rapport* | 21395 (55%) | 8076 (21%) | 3036 (8%) | 6377 (16%) |
| *Thèse* | 239030 (46%) | 141382 (27%) | 61178 (12%) | 77744 (15%) |

Le quatrième tableau présente les résultats inversés du tableau précédent : les grandes catégories d'image sont réparties selon le type de document.

| Classification | Article | Blog | Chapitre | Livre | Mémoires | Rapport | Thèse |
|---|---|---|---|---|---|---|---|
| *Tableaux, textes, logos & artefacts* | 30556 (9%) | 3992 (1%) | 16807 (5%) | 8004 (2%) | 18358 (5%) | 21395 (6%) | 239030 (71%) |
| *Documents* | 3319 (4%) | 10364 (11%) | 4112 (4%) | 1366 (1%) | 10524 (11%) | 3036 (3%) | 61178 (65%) |
| *Graphe, cartes & schémas* | 13521 (7%) | 2593 (1%) | 15201 (8%) | 2450 (1%) | 8915 (5%) | 8076 (4%) | 141382 (74%) |
| *Photographies & peintures* | 6751 (5%) | 15709 (11%) | 16462 (11%) | 4610 (3%) | 15778 (11%) | 6377 (4%) | 77744 (54%) |
| *Total* | 54147 | 32658 | 52582 | 16430 | 53575 | 38884 | 519334 |

Le cinquième tableau donne la répartition dans les 14 principales disciplines recensées dans les métadonnées d'Isidore (à partir des identifiants normalisés de HAL). Sans surprise, les résultats s'accordent aux usages des disciplines et des cultures scientifiques. Les images figuratives (peintures et photographies) sont prédominantes en histoire de l'art (65%) et dans une moindre mesure en archéologie (36%) et en histoire (30%). Inversement, les graphiques et les cartes prédominent en économie et en géographie.

| **Discipline** | **Tableaux, textes, logos et artefacts** | **Graphiques, cartes et schémas** | **Documents** | **Images figuratives** |
|---|---|---|---|---|
| *Anthropologie* | 6576 (48%) | 1790 (13%) | 1166 (9%) | 4028 (30%) |
| *Archéologie* | 15268 (25%) | 17858 (29%) | 6564 (11%) | 22147 (36%) |
| *Histoire de l'art* | 3094 (18%) | 1284 (7%) | 1758 (10%) | 11395 (65%) |
| *Droit* | 10431 (59%) | 2783 (16%) | 3093 (17%) | 1496 (8%) |
| *Économie* | 21490 (56%) | 12105 (32%) | 2894 (8%) | 1875 (5%) |
| *Sciences de l'éducation* | 20657 (45%) | 10418 (23%) | 8075 (18%) | 6535 (14%) |
| *Géographie* | 35876 (43%) | 25896 (31%) | 6518 (8%) | 15757 (19%) |
| *Histoire* | 10723 (26%) | 7656 (18%) | 10237 (24%) | 13241 (32%) |
| *Sciences de l'information et de la communication* | 7794 (38%) | 3951 (19%) | 3922 (19%) | 4736 (23%) |
| *Langues* | 11414 (51%) | 5401 (24%) | 3264 (15%) | 2205 (10%) |
| *LIttérature* | 5048 (41%) | 913 (7%) | 2746 (22%) | 3674 (30%) |
| *Psychologie* | 29291 (64%) | 8820 (19%) | 4202 (9%) | 3209 (7%) |
| *Sciences politiques* | 3790 (52%) | 1346 (18%) | 1040 (14%) | 1129 (15%) |
| *Sociologie* | 11891 (44%) | 4698 (18%) | 4125 (15%) | 6045 (23%) |

Le tableau ci-dessous donne une autre lecture de ces résultats en montrant la répartition par discipline des images relevant de la même classification (autrement dit, les pourcentages se lisent en colonne au lieu de se lire en ligne).

| **Discipline** | **Tableaux, textes, logos et artefacts** | **Graphiques, cartes et schémas** | **Documents** | **Images figuratives** |
|---|---|---|---|---|
| *Anthropologie* | 6576 (3%) | 1790 (2%) | 1166 (2%) | 4028 (4%) |
| *Archéologie* | 15268 (8%) | 17858 (17%) | 6564 (11%) | 22147 (23%) |
| *Histoire de l'art* | 3094 (2%) | 1284 (1%) | 1758 (3%) | 11395 (12%) |
| *Droit* | 10431 (5%) | 2783 (3%) | 3093 (5%) | 1496 (2%) |

| | | | | |
|---|---|---|---|---|
| *Économie* | 21490 (11%) | 12105 (12%) | 2894 (5%) | 1875 (2%) |
| *Sciences de l'éducation* | 20657 (11%) | 10418 (10%) | 8075 (14%) | 6535 (7%) |
| *Géographie* | 35876 (19%) | 25896 (25%) | 6518 (11%) | 15757 (16%) |
| *Histoire* | 10723 (6%) | 7656 (7%) | 10237 (17%) | 13241 (14%) |
| *Sciences de l'information et de la communication* | 7794 (4%) | 3951 (4%) | 3922 (7%) | 4736 (5%) |
| *Langues* | 11414 (6%) | 5401 (5%) | 3264 (5%) | 2205 (2%) |
| *LIttérature* | 5048 (3%) | 913 (1%) | 2746 (5%) | 3674 (4%) |
| *Psychologie* | 29291 (15%) | 8820 (8%) | 4202 (7%) | 3209 (3%) |
| *Sciences politiques* | 3790 (2%) | 1346 (1%) | 1040 (2%) | 1129 (1%) |
| *Sociologie* | 11891 (6%) | 4698 (4%) | 4125 (7%) | 6045 (6%) |

Cette répartition disciplinaire sous-jacente a une incidence sur le corpus visuel de chaque plateforme. Nous avons vu que le taux d'image figuratives sur Cairn est assez faible en comparaison d'OpenEdition et d'autres corpus au format HTML. Le tableau suivant montre la répartition par discipline des documents (principalement des articles) issus de Cairn et d'OpenEdition Journals dans le corpus documentaire. Il en ressort que Cairn héberge davantage d'articles en provenance de disciplines peu illustrées ou recourant principalement à des images non-figuratives (droit, économie, psychologie…).

| **Discipline** | **OpenEdition Journals** | **Cairn** |
|---|---|---|
| *Anthropologie* | 4% (1271) | 4% (476) |
| *Archéologie* | 6% (1793) | 2% (215) |
| *Droit* | 3% (854) | 10% (1265) |
| *Économie* | 3% (835) | 8% (989) |
| *Géographie* | 7% (2202) | 6% (735) |
| *Histoire* | 18% (5652) | 13% (1687) |
| *Histoire de l'art* | 6% (1921) | 4% (476) |
| *Linguistique* | 4% (1305) | 4% (532) |
| *Littérature* | 14% (4241) | 9% (1134) |
| *Psychologie* | 1% (219) | 9% (1124) |
| *Sciences de l'éducation* | 6% (1702) | 6% (816) |
| *Sciences de l'information et de la communication* | 4% (1168) | 5% (616) |
| *Sciences politiques* | 9% (2740) | 9% (1185) |
| *Sociologie* | 16% (4861) | 14% (1796) |

# Métadonnées des images

## Légendes

La légende fait partie intégrante des « bonnes pratiques » de l'édition scientifique. Cet usage devrait théoriquement permettre d'identifier une grande partie du corpus.

Dans le corpus d'OpenEdition, la majorité des images sont légendées avec un titre, un crédit et/ou une description normalisée (il est fréquent que plusieurs de ces champs coexistent). Avec l'extraction automatique des légendes (script create_legend_oe.R) nous avons pu récupérer 86 108 légendes documentant 55 864 images, soit quasiment 60% du corpus visuel d'OpenEdition (qui se monte à 94 850). Les images figuratives ne sont pas significativement plus légendées (25224 images sur 39628 images figuratives au total, soit 63%).

Le nombre réel de légendes est certainement plus élevé : si sur OpenEdition Journals et sur OpenEdition Books, les éditeurs sont attentifs à utiliser les balises normalisées, les pratiques sont beaucoup plus libres sur Hypotheses où l'image peut être simplement décrite par un texte libre. Par conséquent, dans le script d'extraction des légendes nous avons aussi intégré les textes précédant ou succédant à l'illustration même lorsqu'ils n'utilisent pas de balise normalisée.

L'extraction des légendes dans les documents PDF pose plus de difficultés. Elles ne sont en effet pas signalées par des balises normalisées. Le seul indice est "positionnel" : en général les légendes sont situées immédiatement en-dessous de l'image (ou immédiatement au-dessus s'agissant d'un titre) et sont séparées du reste du texte par un saut de ligne. Nous sommes actuellement en train de développer un outil de détection géométrique sur la base de cet usage. Dans le cadre de la phase 3 de l'étude, nous créerons également des outils de détection automatique des données contenues dans les légendes afin d'identifier l'ayant-droit (en s'appuyant sur des méthodes de détection d'entités nommées).

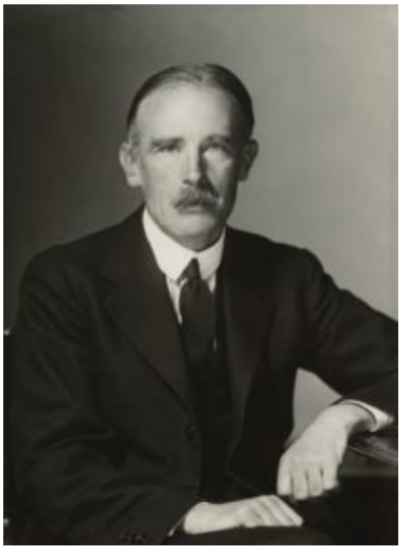
## Métadonnées internes

Même lorsqu'elles ne sont pas documentées par des légendes, les images elles-mêmes peuvent porter des informations documentaires. Il s'agit des métadonnées internes structurées selon plusieurs grands standards : EXIF, IPTC, XMP et IFD0.

Nous avons procédé à l'extraction des métadonnées internes de l'ensemble des objets visuels du corpus à l'aide de la commande `exiftool -a -u -g1 nom_du_fichier`.

Le niveau de détail est très variable. Dans certains cas les images contiennent des liens directs

vers leur provenance et d'autres éléments contextuels très utiles pour statuer sur le statut juridique de l'image (nom de l'auteur, licence). Dans d'autres cas, les métadonnées d'origine ont été écrasées par un traitement intermédiaire et ne sont plus informatives.



```
---- XMP-x ----
XMP Toolkit                     : Adobe XMP Core 5.3-c007 1.136881
---- XMP-xmp ----
Create Date                     : 2006:10:02 10:35:34+01:00
Modify Date                     : 2009:04:25 01:06:53+01:00
Metadata Date                   : 2014:02:10 12:44:37Z
Creator Tool                    : Adobe Photoshop CS3 Macintosh
---- XMP-dc ----
Format                          : image/tiff
Creator                         : National Portrait Gallery London
---- XMP-photoshop ----
Authors Position                : Rights and Images Department
Source                          : National Portrait Gallery, Londo
---- XMP-iptcCore ----
Intellectual Genre              : Portrait
Creator Address                 : St Martin's Place
Creator City                    : London
Creator Postal Code             : WC2H 0HE
Creator Country                 : United Kingdom
Creator Work Email              : rightsandimages@npg.org.uk
Creator Work URL                : www.npg.org.uk
```

*Exemple d'illustration extraite d'OpenEdition avec des données internes très détaillées sur le titulaire des droits et sur le genre de l'image ("Intellectual Genre : Portrait").*

Nous avons défini une liste de champs informatifs en faisant des recherches d'URL sur l'ensemble des données et en éliminant progressivement les champs uniquement techniques. Ces champs proviennent principalement des métadonnées IFD0, XML et IPTC et sont de natures variables : ils peuvent décrire le créateur, le titulaire des droits, le producteur, l'institution ou un commentaire libre.

Sur OpenEdition, 14 058 ont au moins un champ potentiellement informatif et 4 222 disposent d'un champ créateur. Ces métadonnées sont beaucoup moins présentes dans les corpus visuels extraits des PDF. Dans le corpus visuel des thèses, nous ne disposons que de 2 673 images possédant au moins un champ potentiellement informatif (soit moins de 2% du nombre total d'images hors éléments textuels et autres artefacts) et de 3 820 images dans le corpus visuel du CCSD. En théorie, les métadonnées internes ne sont pas supprimées lors de la création du PDF. En pratique, l'usage de logiciels d'édition d'image en parallèle de la création du PDF contribue visiblement à occulter les informations attachées à l'image. Dans Cairn, aucune image n'a de métadonnées informatives, peut-être parce que la plateforme supprime par défaut ce type d'information.

| Collection | Images avec des métadonnées informatives | Proportion des images |
|---|---|---|
| *OpenEdition* | 14058 | 14,80% |
| *Theses* | 2673 | 0,80% |

| | | |
|---|---|---|
| *CCSD* | 3820 | 1% |
| *Cairn* | 0 | 0% |
| *Plateformes PDF* | 449 | 0,70% |
| *Plateformes HTML* | 1085 | 39% |

Les métadonnées internes apparaissent ainsi principalement comme des sources d'information complémentaires. **À la différence de la classification automatique ou de l'extraction des légendes, le volume des informations extraites couvre une partie trop restreinte du corpus pour servir de ressource privilégiée pour les analyses automatisées prévues dans la phase 3 de l'étude. Par contre, dans le cadre de l'annotation d'un échantillon d'images à la phase 2, elles peuvent permettre de documenter les images qui ne sont pas documentées par ailleurs.**

## Estimation du nombre d'images caviardées

Nous avons repéré uniquement deux usages de mentions normalisées d'images caviardées dans le corpus :

- Sur *Persée* les illustrations sont remplacées par la mention standard "illustration non autorisée à la diffusion". Faute d'avoir obtenu communication de ce corpus dans les temps, il sera traité dans le cadre de la phase 5 de l'étude, en même temps que le corpus rétroactif.
- Sur *OpenEdition*, les images manquantes peuvent être signalées par une image portant la mention « Droits numériques non obtenus »[9]. Malheureusement, cette information n'est spécifiée par aucune métadonnée complémentaire, soit sous la forme de légende, soit la forme de données EXIF. À défaut, nous avons intégré un échantillon de ces images dans le modèle de classification sous le label « caviardage ». D'après les résultats de la classification effectuée sur les 94 850 objets visuels d'OpenEdition, il y aurait 159 images caviardées de cette manière.

Dans les échantillons aléatoires issues des images extraites de theses.fr et du CCSD nous n'avons trouvé aucune mention récurrente de caviardage. Les phases ultérieures de l'étude permettront peut-être d'identifier indirectement des masquages occasionnels, par exemple lorsque l'image est remplacée par un grand cadre blanc : le modèle inclut en effet une catégorie « vide », principalement utilisée pour repérer des artefacts de la classification. Ceci dépasse le périmètre de la phase 1.

[9] À titre d'exemple, voir ce chapitre d'ouvrage : https://books.openedition.org/psorbonne/39842

Le croisement des résultats de classification et des données de discipline mises à disposition par Isidore permettent également d'identifier des formes plus diffuses de caviardage. Sur 1 777 articles, chapitres et billets de blogs d'OpenEdition classés dans les études des arts visuels, 1 066 n'ont aucune illustration figurative d'après le modèle (soit près de 60% de l'ensemble) et 819 n'ont aucune image identifiée. Au-delà du caviardage explicite, cette mesure suggère l'usage répandu d'un caviardage par anticipation.

# Phase 2 : analyse documentaire d'un échantillon

## Objectifs

La phase 2 vise à caractériser précisément les images et leurs documents sources, grâce à une analyse documentaire d'un échantillon d'images issues du corpus obtenu à la phase 1.

Il est important de noter que cette analyse documentaire ne constitue pas une analyse juridique du régime des œuvres. Les auteurs ne sauraient être tenus responsables de l'interprétation juridique qui en sera faite.

## Construction de l'échantillon

Rappelons que le corpus final obtenu à la phase 1 est constitué de 110 708 documents comprenant 899 567 objets visuels identifiés par un modèle de deep learning[10]. Ces objets visuels ont notamment comme caractéristique d'avoir une taille supérieure ou égale à 1,5 kilo-octets afin de diminuer une grande partie du bruit (notamment des artefacts de l'extraction automatique).

Le Comité de pilotage #2 du 8 mars 2021 a permis de définir les règles de construction de l'échantillon à partir de ce corpus :

- exclure les images caviardées
- inclure les images figuratives, c'est-à-dire les photographies, dessins et peintures (qui sont au nombre de 174 466 soit 20 % des images)
- inclure les images de la catégorie "documents" (qui sont au nombre de 68 711 soit 8 % des images) qui peuvent être des affiches, des posters, ou des panels d'images
- inclure les images de la catégorie "carte" qui auraient pu être mal attribuées, c'est-à-dire dont la catégorie secondaire relève de l'image figurative avec une probabilité supérieure à 30% (qui sont au nombre de 5 367 soit 0,6 % des images).

[10] Le corpus comprend 487 documents de la plateforme Persée qui restent encore à obtenir. Par conséquent les images de ces documents ne sont pas comptabilisées et seront analysées à la phase 5 de l'étude.

Le seuil de poids des images à 1,5 ko n'a finalement pas été remis en cause : les tests effectués sur des images de 2 ko, 10 ko et 20 ko n'ont pas permis de déterminer de façon consensuelle une limite plus pertinente.

En suivant ces contraintes de départ, nous tirons au sort un échantillon de 1 500 images en respectant une stratification par collection et par catégorie d'image définie plus haut (c'est-à-dire que le nombre d'images d'une collection et d'une catégorie données dans l'échantillon est proportionnel au nombre d'images observé pour cette classe dans le corpus, en s'assurant que chaque classe est représentée par au moins une image). En outre, nous fixons un plafond de 5 images par document afin de varier au maximum les sources.

Ainsi, l'échantillon est le plus représentatif possible et on pourra plus facilement en extrapoler les résultats à l'ensemble du corpus (ce sera l'objet de la phase 4).

Au cours de l'étude nous avons souhaité élargir l'échantillon pour couvrir mieux les collections "Autres corpus en PDF", "Autres corpus en HTML" et "CAIRN", à la mesure de leur hétérogénéité, avec une majoration de 10% (soit 160 images au lieu de 146).

Le tableau ci-après présente les images de départ dans le corpus et, dans la dernière colonne, l'échantillon qui en est tiré (dont, entre parenthèse, le chiffre obtenu après avoir "dopé" légèrement certaines catégories d'images) :

| Collection | Catégorie | Nb d'images | Proportion | Échantillon |
|---|---|---|---|---|
| *OpenEdition* | Cartes incertaines | 775 | 0,31% | 5 |
| | Documents | 16 076 | 6,47% | 97 |
| | Images figuratives | 39 643 | 15,95% | 239 |
| *Theses* | Cartes incertaines | 1 777 | 0,71% | 11 |
| | Documents | 16 713 | 6,72% | 101 |
| | Images figuratives | 43 775 | 17,61% | 264 |
| *CCSD* | Cartes incertaines | 1 986 | 0,80% | 12 |
| | Documents | 29 998 | 12,07% | 181 |
| | Images figuratives | 74 014 | 29,78% | 444 |
| *Cairn* | Cartes incertaines | 18 | 0,01% | 1 |
| | Documents | 1 067 | 0,43% | 6 |

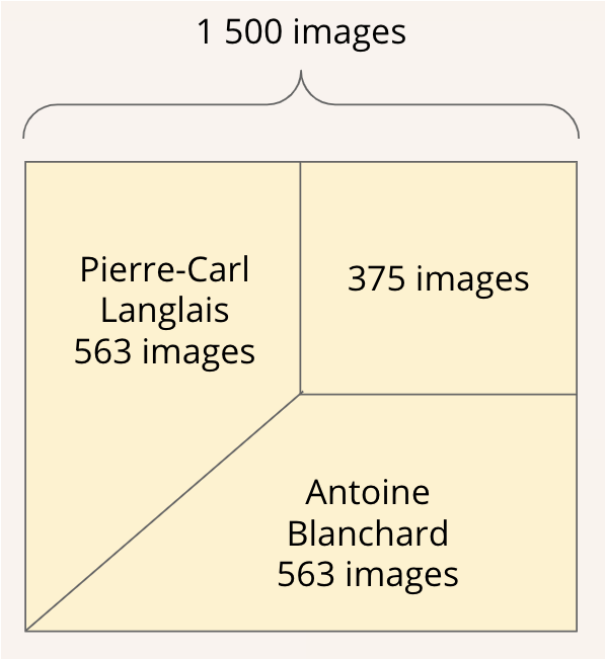
| | | | | |
|---|---|---|---|---|
| | Images figuratives | 2 625 | 1,06% | 16 (18) |
| *Autres corpus en PDF* | Cartes incertaines | 798 | 0,32% | 5 |
| | Documents | 4 438 | 1,79% | 27 (30) |
| | Images figuratives | 13 088 | 5,27% | 79 (87) |
| *Autres corpus en HTML* | Cartes incertaines | 13 | 0,01% | 1 |
| | Documents | 424 | 0,17% | 3 |
| | Images figuratives | 1 336 | 0,54% | 8 (9) |
| **Total** | **—** | **248 564** | **100%** | **1 514** |

**La taille finale de l'échantillon analysé est donc de 1 514 images.**

## Méthode d'analyse

L'analyse documentaire est une analyse au cas par cas des caractéristiques de chaque image. Pour être reproductible, cette analyse doit être soumise à des règles les plus précises possibles, et sujette au moins d'interprétation subjective possible.

Il convient donc de préciser que l'échantillon a été annoté à parts égales par Pierre-Carl Langlais et Antoine Blanchard, avec un quart d'images en commun afin de vérifier l'accord inter-annotateur (les désaccords éventuels étant discutés et rectifiés) :




## Champs de description

Pour chaque image a été renseigné un ensemble de champs descriptifs visant à répondre à une série de questions :

| Nom du champ | Question | Consigne | Sources de référence |
|---|---|---|---|
| Arts_visuels | Est-ce que l'image relève des arts visuels ? | **T** (pour TRUE) si l'œuvre relève des arts visuels : arts plastiques (peinture, sculpture, dessin, installations, gravure, tag/street art, bande dessinée…), photographie, œuvres des arts graphiques et des arts appliqués et du design (notamment les affiches et posters qui relèvent d'un travail de graphisme), créations architecturales (bâtiments, aménagements publics, jardins), dessins d'architecture<br>**F** (pour FALSE) sinon et on | |

| | | | |
|---|---|---|---|
| | | passe à l'image suivante | |
| Susceptible_DA | Est-ce que l'image est susceptible d'être protégée par le droit d'auteur ? | **T** si auteur de l'œuvre (représentée ou de la photographie) vivant, ou mort après 1949<br>**F** si numérisation à l'identique d'une œuvre du domaine public ou photographie non créditée à un photographe d'une œuvre du domaine public, et on passe à l'image suivante<br>**vide** sinon et renseigner la colonne Probabilite_DA | ? recherche de l'auteur dans la base de données de l'ADAGP<br>? si l'auteur n'y est pas, recherche sur Wikipédia |
| Probabilite_DA | Quelle est la probabilité que l'image non attribuée soit protégée par le droit d'auteur ? | **valeur** comprise entre 0 et 1, calculée par année * 0,017 - 32,14, si auteur de l'œuvre (représentée ou de la photographie) indéterminé et œuvre datée entre 1890 et 1950 | |
| Auteurs_DA | L'auteur de l'image est-il un des auteurs du document ? | **T** si auteur de l'image est un auteur du document<br>**F** sinon<br>**vide** si auteur indéterminé | |
| Projet_DA | L'auteur de l'image appartient-il au projet ou à l'équipe qui est à l'origine du document ? | **T** si l'auteur de l'image appartient au projet ou à l'équipe qui est à l'origine du document<br>**vide** sinon | |
| Echec_attribution | Est-ce qu'on a échoué à attribuer précisément l'image ? | **T** si l'œuvre est susceptible d'être couverte par le droit d'auteur mais l'analyse documentaire n'a pas permis d'en déterminer l'auteur | |

| | | | |
|---|---|---|---|
| | | **F** sinon | |
| Sans_but_lucratif | | **T** si document déposé sur une archive ouverte, si pas d'éditeur professionnel (carnet de recherche, colloque et conférence, mémoire, thèse, HDR, rapport), si l'éditeur est une association loi 1901 ou un établissement public ou para-public<br>**F** sinon et on passe à l'image suivante | ? déterminer l'éditeur d'une revue avec le catalogue de la BnF : recherche par nom ou ISSN puis vérifier le champ "Publication"<br>? déterminer le statut juridique d'un éditeur avec le fichier constitué par Caroline Dandurand dans le cadre de sa mission sur la structuration de l'édition scientifique publique<br>? en cas d'échec, déterminer le statut juridique d'un éditeur avec l'Annuaire des entreprises : recherche par nom puis vérifier le champ "Catégorie juridique" |
| Activite_ESR | Est-ce que les auteurs de la publication sont affiliés ou associés à un établissement d'enseignement supérieur et de recherche ? | **T** si, à la date de la première publication, au moins un des auteurs du document (ou un des coordinateurs dans le cas d'un livre collectif) est étudiant ou personnel affilié ou associé à un organisme ou établissement | ? vérification de la mini-bio de l'auteur sur le document<br>? recherche du nom de l'auteur sur ScanR puis dans le champ "Affiliation" à la date de la |

| | | | |
|---|---|---|---|
| | | d'enseignement supérieur et de recherche public français<br>**F** sinon | publication cliquer sur la flèche ? pour aller sur la fiche de l'entité puis "Voir la liste" des identifiants pour vérifier s'il figure dans le RNSR |
| Attributions_multiples | Est-ce que l'image donne lieu à des attributions multiples (en particulier double attribution à l'auteur de l'œuvre photographiée et à l'auteur de la photographie) ? | **T** si la légende disponible sur la page du document ou les métadonnées internes de l'image l'attribuent à plusieurs auteurs correspondant à plusieurs couches de droit<br>**F** sinon | |
| Agence_photographique | Est-ce que l'image est soumise à paiement de droits de mise à disposition et de diffusion de la part d'agences photographique ? | **T** si la légende disponible sur la page du document ou les métadonnées internes de l'image indiquent que la RMN ou autre agence photographique a des droits sur l'image d'une œuvre appartement au domaine public<br>**F** sinon | |
| Licence_libre | Est-ce que l'image est diffusée avec une licence Creative Commons (ou autre licence libre) ? | **T** si la légende disponible sur la page du document ou les métadonnées internes de l'image indiquent qu'elle est diffusée avec une licence Creative Commons ou autre licence libre, ou si le document lui-même est placé sous une licence | |

| | | | |
|---|---|---|---|
| | | libre dans le cas où l'auteur du document est auteur de l'image<br>**F** sinon | |
| Metadonnees_auteur | Qui est l'auteur de l'œuvre représentée et/ou de la photographie ? | **auteur** de l'œuvre représentée et auteur de la photographie, séparés par un tiret bas "_"<br><br>Exemples :<br>_Robert Doisneau<br>Constant Alexandre Famin_Patrice Schmidt | |
| Metadonnees_ayantDroit | Qui est la personne morale revendiquant des droits sur l'œuvre représentée et/ou la photographie ? | personne morale revendiquant des droits sur l'œuvre représentée et personne morale revendiquant des droits sur la photographie, séparés par un tiret bas "_"<br><br>Exemples :<br>Domaine public_RMN<br>Domaine public_ | |
| Commentaire | | **commentaire libre** permettant de justifier un raisonnement ou de relever une caractéristique spécifique de l'image | |

L'interprétation de certains cas limites a été soumise au Comité de pilotage : le rapport ne revient pas sur la façon dont ces cas ont été traités.

## Sources de référence

Les sources de référence sont citées dans la dernière colonne du tableau précédent. Voici quelques précisions les concernant.

Nous avons considéré comme organisme ou établissement d'enseignement supérieur et de recherche public français trois structures absentes du RNSR (Répertoire national des structures de recherche) :

- l'Institut de recherche stratégique de l'Ecole militaire (IRSEM) sous tutelle du Ministère des armées (créé par l'Arrêté du 22 décembre 2015 portant organisation de l'institut de recherche stratégique de l'Ecole militaire)
- le Centre de recherche du château de Versailles, GIP (créé par l'Arrêté du 27 octobre 2006 portant approbation de la convention constitutive d'un groupement d'intérêt public) ayant comme membres le Ministère de la culture et plusieurs EPSCP sous tutelle du MESRI
- l'Agence bibliographique de l'enseignement supérieur (Abes), établissement public à caractère administratif sous tutelle du MESRI (créée par le Décret n° 94-921 du 24 octobre 1994 portant création de l'Agence bibliographique de l'enseignement supérieur).

Concernant le fichier constitué par Caroline Dandurand dans le cadre de sa mission sur la structuration de l'édition scientifique publique, qu'ont été considérés comme éditeurs à but lucratif les :

- GIE (exemple : Quae Éditions)
- SASU (exemple : Presses de l'EHESP)
- SARL (exemple : Presses de Sciences Po)
- SA (exemple : CNRS Éditions)

À l'inverse, sont considérés comme éditeurs à but non lucratif les services propres ou communs d'établissements d'enseignement supérieur et de recherche ainsi que leurs Services d'activités industrielles et commerciales (les SAIC prévus par l'Article L313-1 du Code de la recherche), mais également les fondations et les associations.

## Résultats de l'analyse

### Résultat général

Les principaux champs d'annotations correspondent à un arbre de décision : chaque image passe une série de "tests" successifs. Nous évaluons consécutivement si :

- une image relève des arts visuels
- elle est susceptible d'être protégée par le droit d'auteur
- elle est sous licence libre
- elle est présente dans une publication à but non lucratif
- elle est présente dans une publication d'un auteur affilié à un établissement de l'ESR français
- elle n'est pas une création personnelle de l'auteur
- elle n'est pas une création d'un auteur appartenant au projet ou à l'équipe à l'origine du document.

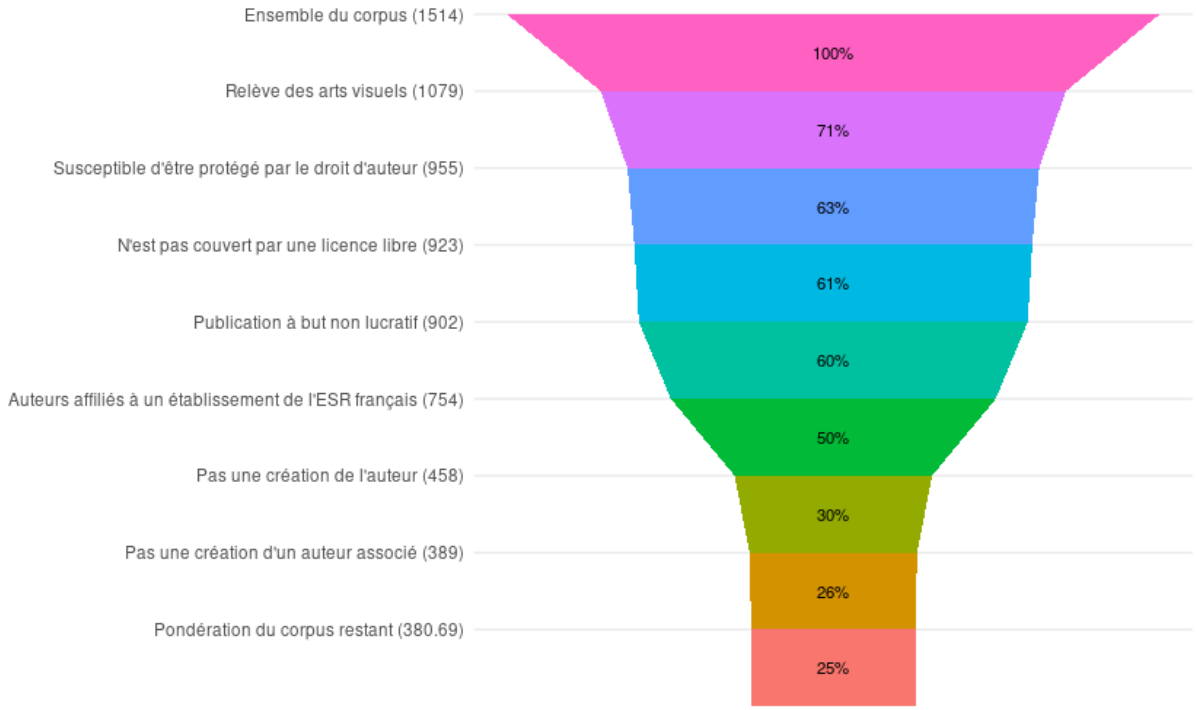
Concrètement, si une image créée par l'auteur de la publication ne relève pas des arts visuels alors elle ne sera pas comptabilisée comme création personnelle de l'auteur.

Le tableau ci-dessous décrit chaque test successif en mentionnant les images conservées ("oui"), les images écartées ("non") et la proportion d'images conservées.

| Critère | Non | Oui | Proportion d'images conservées |
|---|---|---|---|
| Images dans l'échantillon | | 1514 | 100% |
| Relève des arts visuels | 435 | 1079 | 71,27% |
| Susceptible d'être protégé par le droit d'auteur | 124 | 955 | 63,08% |
| N'est pas couvert par une licence libre | 32 | 923 | 60,96% |
| Publication à but non lucratif | 21 | 902 | 59,58% |
| Auteurs affiliés à un établissement de l'ESR français | 148 | 754 | 49,80% |
| Pas une création de l'auteur | 296 | 458 | 30,25% |
| Pas une création d'un auteur appartenant au projet ou à l'équipe | 69 | 389 | 25,69% |
| | | | |
| **Sous-total** | | **389** | **25,69 %** |
| **Sous-total pondéré par la probabilité d'être sous droit d'auteur** | | **380,69** | **25,14 %** |

Les deux sous-totaux reprennent d'une part le nombre total d'images ayant passé tous les tests, et d'autre part une pondération par la probabilité d'être sous droit d'auteur ; en effet, pour les images datées dont l'auteur n'a pas pu être déterminé, nous avons appliqué une règle de calcul de probabilité telle que définie pour le champ Probabilite_DA (chaque probabilité étant comprise entre 0 et 1, le sous-total pondéré est logiquement inférieur).

Ces résultats peuvent être représentés sous la forme d'un entonnoir, montrant l'attrition progressive de l'échantillon après application de chaque critère :

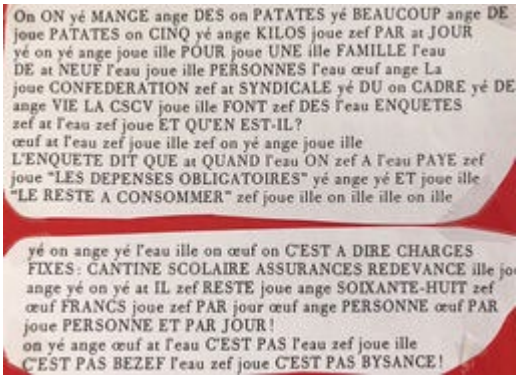



Le tableau ci-dessous donne à titre complémentaire, la proportion des tests négatifs par rapport à l'ensemble de l'échantillon :

| **Critère** | **Nombre d'images** | **Proportion d'images de l'échantillon** |
|---|---|---|
| Image qui ne relève pas des arts visuels | 435 | 28,73 % |
| Image non susceptible d'être protégé par le droit d'auteur | 124 | 8,19 % |
| Image sous licence libre | 32 | 2,11 % |
| Image dans une publication à but lucratif | 21 | 1,39 % |
| Auteurs non affiliés à un établissement de l'ESR français | 148 | 9,78 % |
| Création de l'auteur | 296 | 19,55 % |
| Création d'un auteur appartenant au projet ou à l'équipe | 69 | 4,56 % |
| | | |
| **Sous-total** | **389** | **25,69 %** |
| **Sous-total pondéré par la probabilité d'être sous droit d'auteur** | **380,69** | **25,14 %** |

Ces résultats montrent clairement que certains critères sont plus sélectifs que d'autres. L'appartenance aux arts visuels et le fait d'être produit par l'auteur de la publication sont les facteurs les plus sélectifs, suivis par l'affiliation à un établissement de l'ESR français et l'absence de protection au titre du droit d'auteur.

Voici quelques exemples d'images qui n'entrent pas dans le champ des arts visuels et ont été écartées sur le premier critère :

<table>
<tr>
<td></td>
<td>
</td>
</tr>
<tr>
<td>Extrait de CANTEAUT, Olivier ; MOUFFLET, Jean-François. Les éditions d'actes princiers (xiie-xve siècle) : bilan à l'heure du numérique In : Jean de Berry et l'écrit : Les pratiques documentaires d'un fils de roi de France. Paris : Éditions de la Sorbonne, 2019. http://books.openedition.org/psorbonne/54243</td>
<td>Extrait de Francis Grossmann. "Discours rapporté versus Discours partagé : convergences, différences, problèmes de frontières". Article issu de la conférence donnée au colloque Ci-dit, Bruxelles, 2018; soumis à la revue LE DISCOURS ET LA LANGUE. 2019. https://hal.archives-ouvertes.fr/hal-02005379</td>
</tr>
<tr>
<td>
</td>
<td>$$\max_{F_j} U(c_{j3t}, h_{j3,t+1}) = c_{j3t}^{\alpha} h_{j3,t+1}^{1-\alpha}$$
$$s/c\ y_{j3t} = A_3 h_{j3t}^{\beta} n^{1-\beta}$$
$$c_{j3t} = (1-\overline{\tau}) y_{j3t} - e_{\overline{Fj}}$$
$$h_{j3,t+1} = \theta(1-n).h_{j3t}^{\delta}[(\tau y_{1t})^{a}.(\tau y_{2t})^{b}(\overline{\tau} y_{3t})^{c}]^{\gamma/N} \overline{F}_j^{\gamma}$$</td>
</tr>
<tr>
<td>Extrait de Krystina Marcoux. Une méthodologie unique du spectacle vivant : d'après l'analyse des spectacles de Georges Aperghis et de Thierry De Mey. Université de Lyon, 2019. https://tel.archives-ouvertes.fr/tel-02462701</td>
<td>Extrait de Arestoff Florence, Jacques Jean-François, « Politiques éducatives et évasion fiscale dans les pays en développement », Revue d'économie politique, 2016/6 (Vol. 126), p. 1057-1075. https://www.cairn.info/revue-d-economie-politique-2016-6-page-1057.htm</td>
</tr>
</table>

Les deux tableaux suivants montrent la répartition par discipline à partir des métadonnées de classification fournies par Isidore, lesquelles couvrent un peu plus de la moitié de l'échantillon (820 images sur 1 514) en raison de limitations intrinsèques à Isidore. Notons qu'une même publication peut avoir plusieurs appartenances disciplinaires, ce qui explique que la somme des disciplines (29 images en histoire, 175 en archéologie, 64 en architecture…) dépasse 820 :

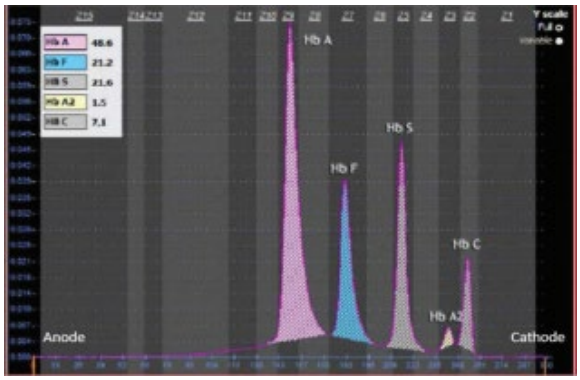
| | Anthropologie (29) | Archéologie (175) | Architecture (64) | Histoire de l'art (83) | Sciences de l'éducation (72) | Géographie (114) |
|---|---|---|---|---|---|---|
| Hors art visuel | 4 (13.79%) | 30 (17,14%) | 9 (14.06%) | 3 (3.61%) | 30 (41.67%) | 29 (25.44%) |
| Hors droit d'auteur | 0 (0%) | 4 (2.29%) | 7 (10.94%) | 17 (20.48%) | 2 (2.78%) | 1 (0.88%) |
| Licence libre | 1 (3.45%) | 7 (4%) | **4 (6.25%)** | 0 (0%) | 0 (0%) | 2 (1.75%) |
| Publication à but lucratif | **2 (6.9%)** | 4 (2.29%) | 0 (0%) | 0 (0%) | 0 (0%) | 1 (0.88%) |
| Hors ESR français | 2 (6.9%) | **50 (28.57%)** | 6 (9.38%) | 10 (12.05%) | 5 (6.94%) | 7 (6.14%) |
| Création de l'auteur | 11 (37.93%) | 43 (24.57%) | 13 (20.31%) | 12 (14.46%) | 20 (27.78%) | **41 (35.96%)** |
| Création d'un auteur appartenant au projet ou à l'équipe | 1 (3.45%) | 19 (10.86%) | 0 (0%) | 3 (3.61%) | 3 (4.17%) | 8 (7.02%) |
| Corpus restant | 8 (27.59%) | 18 (10.29%) | **25 (39.06%)** | **38 (45.78%)** | 12 (16.67%) | 25 (21.93%) |
| **Total** | 29 | 175 | 64 | 83 | 72 | 114 |

| | Histoire (120) | Linguistique (48) | Info-com (32) | Littératures (44) | Psychologie (34) | Sociologie (57) |
|---|---|---|---|---|---|---|
| Hors art visuel | 34 (28.33%) | 15 (31.25%) | 15 (44.12%) | 12 (27.27%) | **16 (47.06%)** | 18 (31.58%) |
| Hors droit d'auteur | **29 (24.17%)** | 5 (10.42%) | 0 (0%) | 6 (13.64%) | 1 (2.94%) | 1 (1.75%) |
| Licence libre | 1 (0.83%) | 1 (2.08%) | 0 (0%) | 1 (2.27%) | 0 (0%) | 2 (3.51%) |
| Publication à but lucratif | 0 (0%) | 0 (0%) | 0 (0%) | 1 (2.27%) | 0 (0%) | 1 (1.75%) |
| Hors ESR français | 22 (18.33%) | 3 (6.25%) | 0 (0%) | 3 (6.82%) | 0 (0%) | 1 (1.75%) |
| Création de l'auteur | 11 (9.17%) | 4 (8.33%) | 7 (20.59%) | 7 (15.91%) | 10 (29.41%) | 10 (17.54%) |
| Création d'un auteur | 2 (1.67%) | **6 (12.5%)** | 0 (0%) | 2 (4.55%) | 2 (5.88%) | 1 (1.75%) |

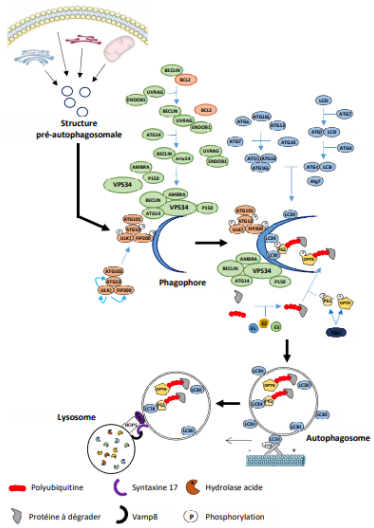
| | | | | | | |
|---|---|---|---|---|---|---|
| appartenant au projet ou à l'équipe | | | | | | |
| Corpus restant | 21 (17.5%) | 14 (29.17%) | 12 (35.29%) | 12 (27.27%) | 5 (14.71%) | 23 (40.35%) |
| **Total** | 120 | 48 | 32 | 44 | 34 | 57 |

Nous signalons en gras le résultat le plus élevé pour chaque critère, c'est-à-dire la discipline pour laquelle le critère est le plus sélectif (en pourcentage). Ainsi, l'attrition des images issues de publications en histoire de l'art et en architecture est beaucoup plus faible avec respectivement environ 46 % et 39 % d'images conservées.

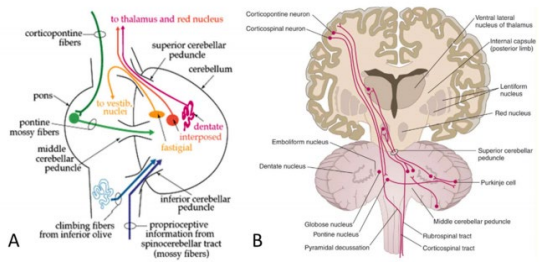
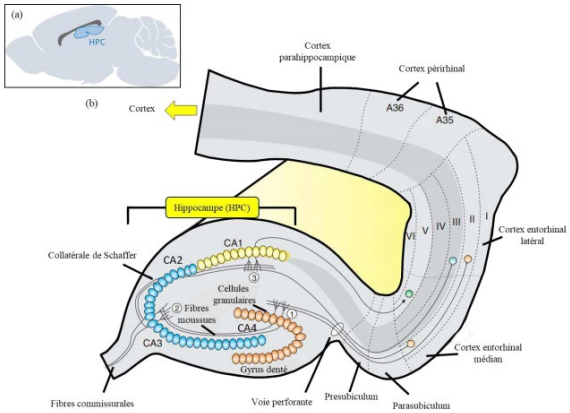
Certaines de ces observations peuvent s'expliquer par la nature même des disciplines et des documents qu'elles publient. Ainsi la psychologie utilise proportionnellement moins d'images relevant des arts visuels en raison de la présence, dans son versant biologique, de représentations graphiques ou schématiques :




Extrait de Fatou Gueye. *Drépanocytose et polymorphismes génétiques : épidémiologie, prédiction de gravité et stress-oxydant*. Université de Lyon; Université Cheikh Anta Diop (Dakar), 2019. https://tel.archives-ouvertes.fr/tel-02310645

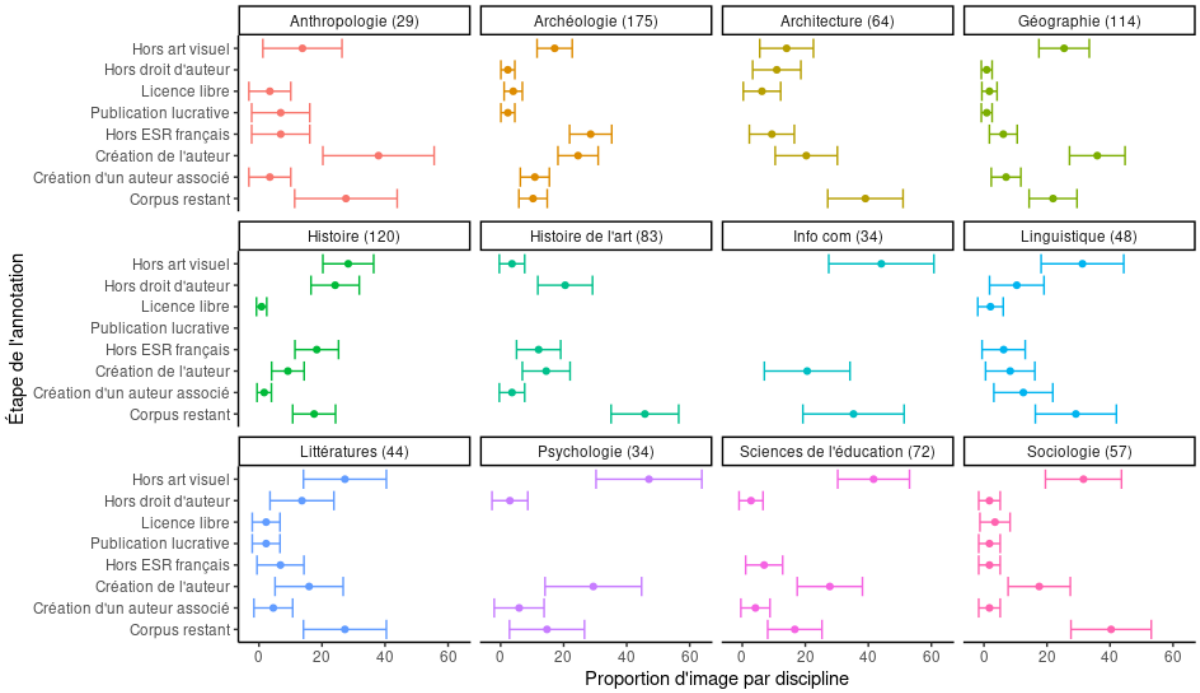



Extrait de Robin Waegaert. *Etude du continuum mécanistique et physiopathologique entre la Sclérose Latérale Amyotrophique et la Démence FrontoTemporale*, Thèse. Université de Strasbourg. 2019. http://www.theses.fr/2019STRAJ110/document

| | |
|---|---|
|  |  |
| Extrait de Marlène Rasschaert. *Capture cérébrale de chélates de gadolinium : imagerie multimodale et analyse des conséquences neurotoxicologiques*. Université Paris Saclay, 2019. https://tel.archives-ouvertes.fr/tel-02294184 | Extrait de Céline Duffau. *Implication de la connectivité anatomo-fonctionnelle des régions du cortex frontal dans la formation et la consolidation de la mémoire associative olfactive chez le rongeur*. Université de Bordeaux, 2019. https://tel.archives-ouvertes.fr/tel-02869584 |

L'histoire utilise proportionnellement plus d'images du domaine public car elle mobilise des corpus plus anciens. Les documents en archéologie sont plus souvent écrits par des auteurs non affiliés à des établissements d'enseignement supérieur et de recherche, du fait que de nombreuses fouilles sont confiées à des associations (comme le Centre d'études et de recherches archéologiques du Morbihan) ou des bureaux d'études archéologiques (comme Éveha) qui rédigent ensuite le rapport de fouille. Les chercheurs en géographie utilisent plus souvent leurs propres images car ce sont typiquement des photographies prises sur le terrain, dans les pays ou régions étudiés. Les chercheurs en linguistique, et pas loin derrière en archéologie, utilisent plus souvent des images prises par des collaborateurs appartenant au projet ou à l'équipe, en particulier s'agissant des techniciens des équipes de fouille.

La visualisation ci-dessous permet de mieux qualifier ces résultats en tenant compte d'une marge d'erreur de 95 % (calculée selon les méthodes des sondages). Chaque point correspond à la proportion estimée d'image par discipline pour chaque étape de l'annotation. La barre représente la distance entre la valeur minimale et la valeur maximale de la marge d'erreur. Par exemple, 25 % des images en histoire ne sont pas susceptibles d'être protégées par le droit d'auteur : avec une marge d'erreur de 7,5 %, l'estimation se situe dans une fourchette entre 16,5 % et 31,5 %.




Comme dans un sondage classique, la marge d'erreur dépend notamment du nombre d'images présentes pour chaque discipline : plus l'échantillon est important et plus la marge d'erreur est faible. En psychologie où nous n'avons que 34 images, la part d'images hors arts visuels peut être estimée entre 30 % et 64 % (marge d'erreur à 16,8 %). En archéologie où nous avons 175 images, la part d'images hors arts visuels tient dans une fourchette beaucoup plus réduite entre 11,56 % et 22,72 % (marge d'erreur à 5,5 %). La visualisation donne ainsi un aperçu du degré de certitude et d'incertitude des résultats.

Cette représentation statistique sera de nouveau mobilisée pour reporter les résultats de l'échantillon à l'ensemble du corpus lors de la phase 4 de l'étude et pour construire une méthode d'extrapolation de cette distribution à d'autres années que 2019.

Il est possible également de ventiler les images dans le champ de la mesure par collection :

| Collection | Échantillon | Nb d'images ayant passé tous les tests | Nb d'images ayant passé tous les tests pondéré par la probabilité d'être sous droit d'auteur | Proportion |
|---|---|---|---|---|
| *Open Edition* | 341 | 89 | 87.2 | 26% |
| *Theses* | 376 | 89 | 88.3 | 24% |
| *CCSD* | 637 | 195 | 189 | 31% |
| *Cairn* | 25 | 2 | 2 | 8% |

| | | | | |
|---|---|---|---|---|
| *Autres corpus HTML* | 13 | 4 | 4 | 30% |
| *Autres corpus PDF* | 122 | 10 | 10 | 8% |

## Résultats complémentaires

D'autres résultats peuvent être présentés pour éclairer de diverses manières l'échantillon.

**69 images sur 1 514 sont produites par une personne participant au projet**. Cela recouvre des situations très diverses liées à des productions scientifiques collaboratives (photographie d'un technicien, d'un collègue du laboratoire ou d'un stagiaire), assez fréquentes en archéologie (Inrap) ou en océanographie (Ifremer). Des publications plus individuelles (thèses, articles) peuvent également s'appuyer sur des images réalisées par des proches, souvent alors mentionnées explicitement dans un texte de remerciements. Ces usages suggèrent que l'auteur aurait obtenu informellement les droits de l'image avant de l'intégrer dans la publication.

**161 images sur 1 514 sont, au moins partiellement, dans le domaine public** (absence de droit d'auteur et/ou mention de "domaine public" dans les métadonnées des ayant droit). En leur sein, 13 images sont encore protégées par des ayants-droit externes à la publication scientifique, ce qui inclut principalement des institutions patrimoniales (le Musée du Louvre, la RMN, l'Institut Français de Pondichéry). À cet ensemble s'ajoutent **21 images sans attribution datées entre 1890 et 1950** qui relèvent potentiellement du domaine public. Nous avons appliqué dans ce cas la formule de calcul fonction de la date de publication recommandé par le Comité de pilotage (voir définition du champ Probabilite_DA).
Nous avons identifié au moins un cas où l'image a été attribuée à tort au domaine public par l'auteur de la publication, sans doute en raison du fait que la photographie est disponible sur Gallica au titre d'un fonds acquis par la BnF :

***En attendant Godot*, de Samuel Beckett, mise en scène de Otomar Krejca, Festival d'Avignon, 1978.**



Phot. Fernand Michaud (Domaine public) source : Gallica BnF

Extrait de Philippe Le Pape, « Vie d'Unimarc en temps de transition bibliographique », *Arabesques* [En ligne], 87 | 2017, mis en ligne le 01 décembre 2019, https://publications-prairial.fr/arabesques/index.php?id=389

**32 images sur 1 514 sont sous licence libre**. Cet ensemble inclut également les créations des auteurs de la publication dans le cas où celle-ci est intégralement placée sous une licence libre, ce que permettent notamment les plateformes OpenEdition et CCSD. Les licences les plus fréquentes sont les Creative Commons (CC-BY-SA, CC-BY-NC).

**Ce nombre est sous-estimé car certaines licences libres sont omises par les auteurs des documents.** C'est le cas de cette photographie de Lilyana Mavrodinova (1932-2016) issue du portail Byzart.eu, où elle est mise à disposition sous licence CC-BY-ND :



culture or accepted the new trends of the Western European culture. The presented pictures reveal a large part of the religious buildings in Bulgaria – temples from the 1st century AD, rock-hewn monasteries, medieval churches and churches of 19th century, the most flourishing period in the construction of Christian buildings. Some of the buildings do not exist anymore. The icon painting heritage also has a vast historical scope (11th

Fig. 2. Church of St Nicholas, Melnik village, 13th century. Institute of Art Studies Archives – Lilyana Mavrodinova Archive.

Extrait de Isabella Baldini, Giulia Marsili, Claudia Lamanna, Lucia Maria Orlandi. *The Silk and the Blood*. Images of Authority in Byzantine Art and Archaeology (Bologna, February 15th, 2019). Inauguration of the digital exhibition and proceedings of the final meeting of "Byzart - Byzantine Art and Archaeology on Europeana" project. 2019, https://halshs.archives-ouvertes.fr/halshs-02906310

C'est le cas également pour cette photographie issue d'une collection de référence utilisée en sciences du langage. En effet, la série Stimulus Pictures Series for Positional Verbs (PSPV) a été développée au Max Planck Institute sous licence CC-BY-NC-SA, ce qui n'apparaît pas dans la légende ou les métadonnées internes :



FIG. 4.49 : Ficelle et table (PSPV_41)

Extrait de Jhonnatan Rangel. *Variations linguistiques et langue en danger. Le cas du numte ?oote ou zoque ayapaneco dans l'état de Tabasco, Mexique*. Institut National des Langues et Civilisations Orientales (INALCO), 2019, https://tel.archives-ouvertes.fr/tel-02989501

Seulement **9 images sur 1 514 relèvent d'une agence photographique**. Les situations sont très diverses : photographies de stock (Getty Image, Clipart, Pixel Shot…), agences de presse (AFP, Isopix) ou institutions patrimoniales (The Trustees of the British Museum, WienMuseum, Bridgeman Images, RMN…). Deux illustrations ont pu être identifiée en provenance des collections de la RMN : une photographie de la Forêt de Fontainebleau de 1870 utilisée dans un article consacré à Flaubert[11] et une photographie d'une terre cuite dans un article consacré à l'œuvre de Bernard Palissy dont l'autrice n'est pas chercheuse mais conservatrice du patrimoine au musée national de la Renaissance – château d'Écouen[12]. Cet usage très faible suggère que les fonds de la RMN restent sous-utilisés dans les publications contemporaines en libre accès. La phase n°5 de l'étude, consacrée aux revues numérisées par Persée et leurs illustrations caviardées une fois mises en ligne, permettra d'évaluer plus justement l'ampleur des reproductions des illustrations de la RMN dans les publications en sciences humaines et sociales.

Enfin, **15 images sur 1 514 ont des attributions multiples**. Cela recouvre à nouveau des cas très variables : documents numérisés avec attribution de l'institution patrimoniale et/ou du photographe, reproduction d'une création architecturale ou d'autres objets en trois dimensions, photographie provenant d'une banque d'image avec un auteur identifié…

Le croisement réalisé par l'ADAGP avec le répertoire d'auteurs qu'elle gère, mais aussi celui des autres sociétés d'auteurs, a permis de mettre en évidence :

- 4 auteurs membres de la SAIF ;
- 29 auteurs membres de l'ADAGP (ou de sociétés représentées par l'ADAGP comme Abacapress, Bildkunst, Getty Images France, Visda) ;
- 29 auteurs membres d'autres sociétés d'auteurs (ALCS, ASCAP, BMI, DILIA, SACD, SCAM, SGAE, et SOGEM).

## Commentaires

**L'annotation des images de l'échantillon a mis en évidence que les pratiques d'illustration sont variées**, en particulier en ce qui concerne la photographie qui est utilisée comme matériau brut (témoin du phénomène étudié par l'auteur) mais aussi comme agrément (pour enjoliver un

[11] Sylvie Giraud, « Dans les pas de Flaubert en forêt de Fontainebleau », *Flaubert* [En ligne], mis en ligne le 15 mai 2019 http://journals.openedition.org/flaubert/3467

[12] Aurélie Gerbier, « Du modèle au tirage : le moulage dans l'œuvre de Bernard Palissy », Technè [En ligne], 47 | 2019, mis en ligne le 01 juin 2020 http://journals.openedition.org/techne/1448

billet sur un carnet de recherche Hypothèses.org…), ou comme illustration (pour donner à voir un protocole de recherche ou une situation expérimentale où l'auteur peut même se mettre en scène…). Les photographies peuvent aussi être des "stimulus expérimentaux" pour la recherche, par exemple en sciences du langage : dans ce cas les auteurs utilisent des collections de référence, dont les conditions de réutilisation ne sont pas toujours explicites.

**L'archéologie représente une part importante du corpus**, riche en images particulièrement en ce qui concerne les rapports de fouille (disponibles sur le portail HAL de l'Inrap et sur l'archive ouverte du Service de recherche archéologique de Bretagne) avec leurs nombreuses photos de fouille et d'objets découverts.

Aussi, il faut indiquer que le corpus construit à partir du portail Isidore comprend des **publications en science, technologie et médecine** du fait notamment de l'indexation de Thèses.fr, des archives ouvertes de l'Ifremer et de la plateforme I-Revues de l'Inist du CNRS.

**Le Ministère de l'enseignement supérieur, de la recherche et de l'innovation, et les établissements dont il est tutelle, représentent l'essentiel des affiliations des auteurs.** Le Ministère de la culture apparaît à travers les écoles d'architecture (dont les étudiants auto-archivent massivement leurs mémoires de fin d'étude), l'Inrap, et le GIP Centre de recherche du château de Versailles. Les écoles d'art n'apparaissent pas dans l'échantillon. D'autres auteurs qui sont dans le champ du Ministère de la culture n'ont pas été comptabilisés car n'exerçant pas leur activité au titre de la recherche ou de l'enseignement : ils sont rattachés à la Direction générale des patrimoines, à la DRAC de Bretagne ou encore au Musée national de la Renaissance – château d'Écouen.

Plus inattendu, **un petit nombre d'images sont en fait les photos de profil des auteurs** qui apparaissent dans leur mini-bio (sur les plateformes I-Revues et Cairn).

# Phase 3 : test d'une méthode automatique reproduisant les résultats de l'analyse manuelle sur l'échantillon

## Objectifs

La phase 3 vise à développer une méthode (ou plutôt des méthodes) automatique pour reproduire les résultats de la phase 2 (la "vérité terrain"). **Elle revêt une forte dimension expérimentale, pour laquelle ont été mobilisées de nombreuses techniques actuellement**

**développées dans le champ des humanités numériques et du traitement du texte et du document.**

Il convient de noter que ce n'est pas cette méthode qui va permettre d'estimer, dans les prochaines semaines, le nombre d'images dans le champ de la mesure. En effet, elle n'est pas "prête à l'emploi" et doit répondre à des besoins à plus long terme, selon plusieurs cas d'usage possibles :

- identifier les auteurs et ayants droit afin d'aider à la réparation des droits par les organismes de gestion collective ;
- étudier l'impact de la mesure en comparant les pratiques de publication avant et après son entrée en vigueur, en ce qui concerne par exemple les reprises d'images de la Réunion des Musées nationaux (RMN) ;
- réajuster l'estimation statistique année après année, en bornant l'extrapolation statistique (fourchette) et en identifiant les nécessaires mises à jour de la formule de calcul.

## Échantillon test

L'échantillon construit à la phase 2 a servi à la fois de corpus d'entraînement et de corpus-test pour les méthodes automatiques de la phase 3, c'est-à-dire que celles-ci sont confrontées en continu aux annotations manuelles générées à la phase 2 afin d'ajuster au mieux leur performance et d'obtenir la meilleure adéquation possible entre l'annotation manuelle et l'annotation automatique.

Rappelons que **la taille de cet échantillon, décrit dans le rapport de phase 2, est de 1 514 images**.

## Classification des images

**L'échantillon est déjà partiellement le produit d'une méthode automatique qui a fait ses preuves : la classification supervisée des images** a permis d'écarter un grand nombre de productions visuelles qui ne relevaient pas de la mesure (graphes, artefacts, schémas…).

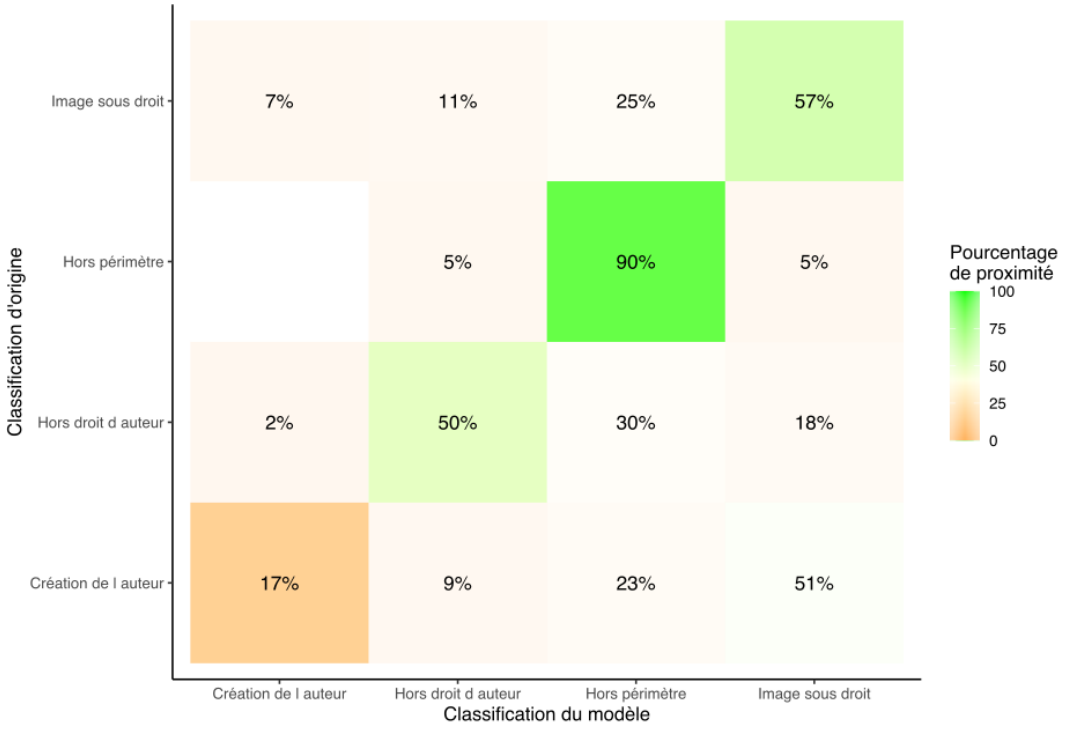
**Nous avons mobilisé à nouveau cette technique à partir des annotations manuelles de la phase 2.** L'observation qualitative de l'échantillon a mis en évidence des régularités visuelles dans deux champs spécifiques de l'annotation : les images ne relevant pas des arts visuels (dites "hors périmètre") et les images patrimoniales dans le domaine public. Cela suggérait qu'il existait encore une marge de manœuvre pour améliorer la pertinence de la modélisation.

Le nouveau modèle a été entraîné sur les quatre catégories les plus importantes du corpus : les images ne relevant pas des arts visuels, les images du domaine public, les créations de l'auteur

et les images sous droit non créées par l'auteur du texte. Incidemment ces catégories semblaient avoir plus de chance de présenter des caractéristiques formelles spécifiques.

Le corpus d'entraînement comprend une sélection aléatoire de 70% de l'ensemble des images annotées. Le modèle a ensuite été expérimenté sur un corpus-test comprenant les 30% d'images restantes. Dans l'ensemble, le modèle a un taux de succès de 68%. Ces résultats sont d'emblée plus faibles que ce que nous avons pu obtenir pour le modèle général pendant la phase 1 de l'étude : les grandes catégories (graphique, document, image figurative…) avaient un taux d'erreur de l'ordre de 5%.

**Le graphique ci-dessous met en évidence de grandes disparités selon les catégories.** Il peut se lire ainsi : les images annotées à la phase 2 comme des créations de l'auteur (rangée du bas) ont été classées à 17% comme des créations de l'auteur, à 9% comme des images du domaine public (hors droit d'auteur), à 23% comme des images ne relevant pas des arts visuels (hors périmètre) et à 51% comme des images sous droit. La somme des pourcentages en ligne vaut toujours 100%.




Résultats du modèle (axe horizontal)
en regard de la classification d'origine (axe vertical)

**Le taux de reconnaissance des images hors périmètre est de loin le plus élevé, avec 90% de succès,** ce qui s'explique probablement par la part élevée d'images non-figuratives. Le modèle est peu performant pour les images sous droits (57%) et les images hors droit d'auteur (50%) et totalement inopérant pour les créations de l'auteur (17%). Manifestement, les images produites

par l'auteur d'un document ne peuvent pas être facilement distinguées (au niveau de leurs caractéristiques visuelles) des images produites par des tiers.

**Ces résultats montrent que la classification des images ne peut pas se substituer à l'annotation humaine. Elle peut par contre contribuer à affiner le corpus à annoter en écartant des images non pertinentes qui avaient passé le filtre de la classification générale mis en œuvre à la phase 1.**

## Analyses automatisée des légendes

L'apparence de l'image n'est pas le seul élément interne au document permettant de juger si une image relève ou non du champ de la mesure. **Les images incluses dans les publications scientifiques sont typiquement accompagnées de légendes décrivant leur titre, leur provenance et éventuellement leurs auteurs et ayants droit** — même si l'annotation manuelle de la phase 2 a permis de constater que cette règle était très souvent ignorée.

### Extraction des légendes des documents au format PDF

Dans les sources publiées en ligne, les légendes sont généralement spécifiées par une syntaxe normalisée. Les revues d'OpenEdition Journals utilisent ainsi tout un jeu de classes prédéfinies qui distinguent le titre (*titreillustration*), le crédit (*créditillustration*) ou la légende au sens strict (*legendeillustration*).

**L'extraction des légendes dans les documents PDF pose plus de difficultés.** Elles ne sont en effet pas signalées par des balises normalisées. Le seul indice est "positionnel" : en général les légendes sont situées immédiatement en dessous de l'image (ou immédiatement au-dessus s'agissant d'un titre) et sont séparées du reste du texte par un saut de ligne. Plusieurs approches ont déjà été expérimentées lors de la préparation de la phase 2 afin d'extraire automatiquement le texte des légendes : les résultats avaient été très mitigés. Même lorsque la légende était correctement reconnue, elle était fréquemment tronquée, en l'absence d'une définition vraiment normée du paragraphe dans les documents PDF.

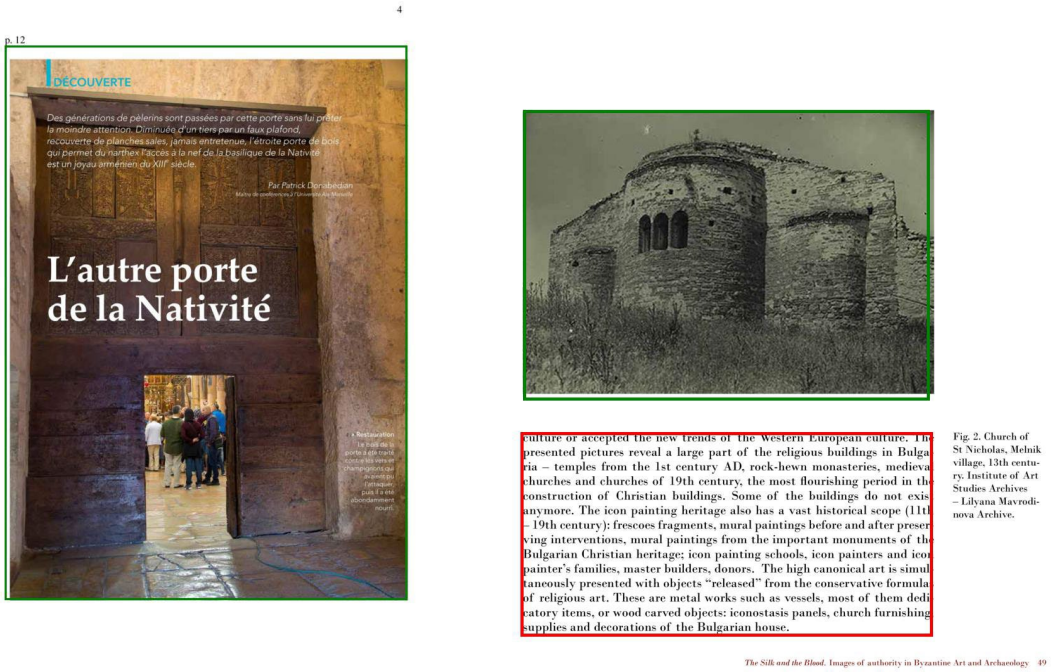
**Dans le cadre de la phase 3 de l'étude, nous avons pu tirer parti d'un nouvel outil de segmentation éditoriale en deep learning : *layoutparser*.** Comme pour notre outil de classification d'image, les modèles mis à disposition par *layoutparser* ont été entraînés à reconnaître la structure des publications scientifiques à partir d'un grand corpus annoté de plusieurs milliers de pages. Le modèle que nous avons utilisé, *PubLayNet*[13], n'identifie pas les

[13] https://github.com/ibm-aur-nlp/PubLayNet

légendes mais quatre grands types d'objets : les titres, les paragraphes, les tableaux et les images. Les deux illustrations ci-dessous donnent un aperçu des résultats de la segmentation : chaque objet éditorial est encadré avec une coloration différenciée (rouge pour le texte, vert pour les images, jaune pour les tableaux et bleu pour les titres).



Le modèle reste cependant perfectible : il a été élaboré à partir du format classique de l'article anglo-saxon en sciences et technologies (avec notamment un découpage de rigueur en deux colonnes). Malgré une mise en page complexe croisant graphes, titres, données et tableaux, le modèle donne ainsi de bons résultats sur la troisième page de la galerie précédente. Inversement, certaines formes de légendes ne sont pas identifiées par le modèle comme les légendes incrustées dans l'image ou les légendes disposées dans la marge, probablement parce qu'elles ne correspondent pas aux usages éditoriaux du corpus d'entraînement.




**Les avancées et les limites de *layoutparser* mettent en évidence l'intérêt de développer un**

**modèle de segmentation adapté aux documents scientifiques français. L'enjeu de ce travail va bien au-delà de la reconnaissance des légendes pour identifier l'ayant droit d'une illustration.** La segmentation rend le document PDF "découvrable" au même titre que les documents web et améliore leur indexation dans un moteur de recherche spécialisé comme Isidore.

Dans le cadre de nos travaux expérimentaux, nous avons limité l'extraction des légendes aux deux principaux corpus au format PDF : Theses.fr et les archives ouvertes du CCSD. Avec les légendes déjà extraites d'OpenEdition, nous avons pu constituer un sous-corpus de 733 images légendées sur 1 353 images relevant de ces trois collections (soit 54%).

L'absence de légende pour 46% de l'échantillon s'explique par une combinaison de facteurs :

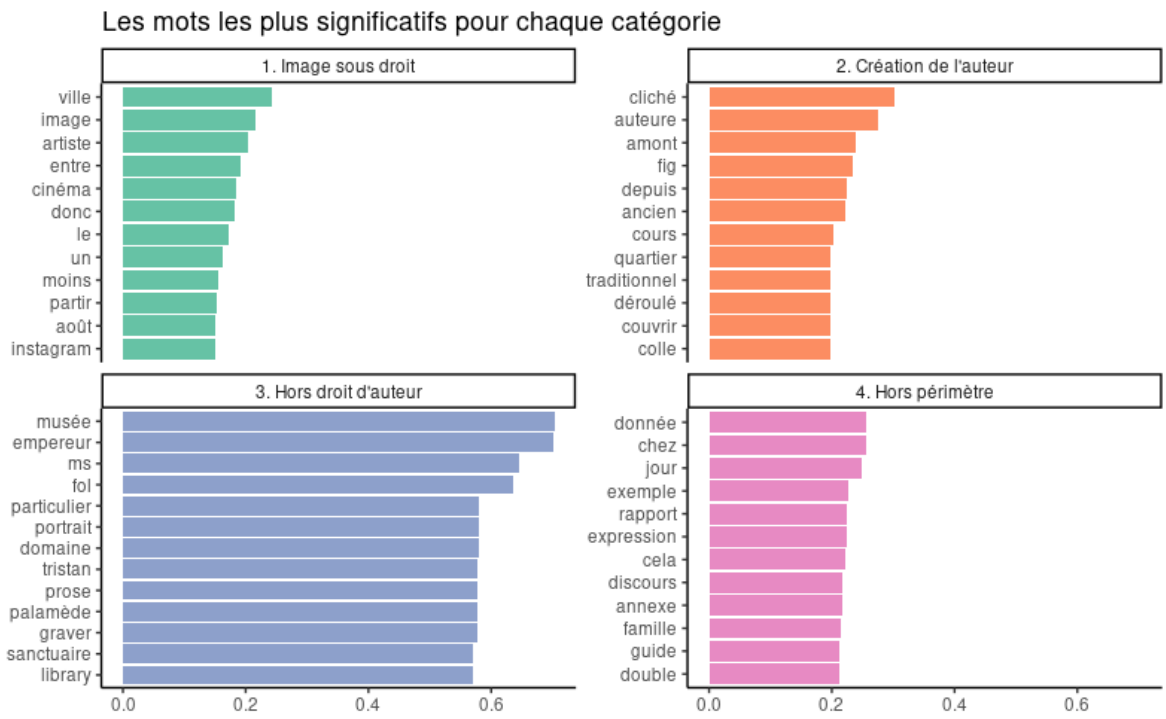
- certaines images n'ont tout simplement pas de légende, par exemple lorsque l'image est en réalité une production de l'auteur ;
- la légende est intégrée dans le texte du document et ne constitue pas un objet éditorial autonome. C'est notamment le cas dans des publications plus informelles comme les carnets de recherche d'Hypotheses ou les mémoires d'étudiants déposés sur la plateforme Dumas ;
- la légende correspond à un format non reconnu par *layoutparser* (légende sur le côté, légende incrustée, etc.). Ce dernier cas de figure pourrait être partiellement résolu avec la création d'un nouveau modèle de segmentation éditoriale adapté aux publications scientifiques françaises en sciences humaines et sociales.

### Classification des légendes

L'observation qualitative du corpus a mis en évidence l'existence de marqueurs plus ou moins normalisés du statut légal des images dans le texte des légendes. Par exemple, 45 légendes utilisent le sigle du copyright ("©") qui n'a pas de valeur juridique en français mais signale généralement une attribution à un ou plusieurs ayants droit.

**Nous avons créé un modèle lexical en SVM (*support vector machine*) qui tente de prédire le statut légal de l'image à partir du contenu textuel des légendes[14]. Concrètement le modèle parcourt l'échantillon déjà annoté et assigne une probabilité de classification associée à chaque mot.** L'image ci-dessous montre les termes utilisés dans les légendes qui sont les plus corrélés à chaque grande catégorie :

[14] Modèle "classification_legende.rda" déposé dans le dossier "modeles" du dépôt de codes sources : https://gitlab.huma-num.fr/planglais/images-usages-isidore/-/tree/main/modeles




Mots contribuant le plus le plus à chaque catégorie
dans le modèle de classification supervisé des légendes.

Le graphique peut se lire ainsi : le mot "ville" a une probabilité d'environ 30% de légender une image sous droit, et le mot "musée" a une probabilité d'environ 70% de légender une image du domaine public.

En apparence, cette distribution conforte l'hypothèse initiale. Les images hors périmètres correspondent à des représentations non figuratives ("données", "rapport", "annexe"). Les images hors droit d'auteur relèvent principalement de collections patrimoniales ("musée", "ms" pour manuscrit, "graver"). Les créations de l'auteur sont des "clichés" pris par les "auteurs" et peuvent notamment porter sur des espaces géographiques que l'auteur a exploré. Enfin les images sous droit d'auteur sont davantage définies en creux : elles peuvent être attribuées à des "artistes"

**Les résultats du modèle sur le corpus-test sont cependant limités avec seulement 47% de succès.** Dans le détail, le modèle est plutôt complémentaire du modèle précédent de classification par image : il est plus efficace sur la catégorie la plus "faible", celle des images créés par l'auteur (67% de classification correcte) :




Résultats du modèle (axe horizontal)
en regard de la classification d'origine (axe vertical)

**D'autres tests menés avec des modèles plus sophistiqués en *deep learning* (*transformers*) n'ont pas donné de résultats plus concluants.** Soit l'échantillon est trop restreint, soit les indications contenues dans les légendes ne suffisent pas pour prédire le statut légal d'une image.

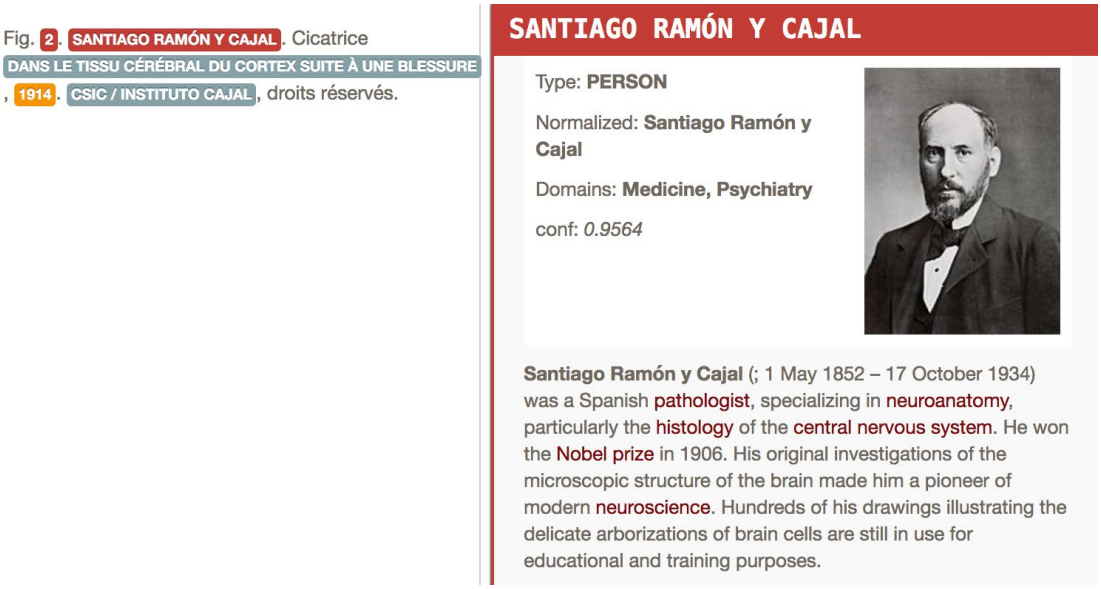
## Extraction des entités nommées

Les méthodes utilisées jusqu'à présent permettent seulement d'inférer le statut probable de l'image à partir d'indices morphologiques contenus dans l'image elle-même ou dans sa légende.

**La détection d'entités nommées rend possible l'identification des ayants droit. Cette technique repose sur l'extraction contextuelle d'entités (généralement des noms d'organisations, de personnes ou de lieux) à partir de la construction syntaxique de la phrase.** Dans notre échantillon, les noms de personne et d'organisation constituent des ayants droit potentiels.

Nos tests ont été menés d'abord avec Spacy, une bibliothèque logicielle en Python spécialisée dans le traitement automatique du langage naturel puis sur l'application EntityFishing mise à disposition par Huma-Num. Par rapport à Spacy, EntityFishing ne permet pas seulement d'identifier les entités mais aussi de les désambiguïser et de les "lier" à une base de référence, Wikidata. À chaque nom de personne, d'organisation ou de lieu potentiellement présent sur Wikidata, EntityFishing associe un identifiant normalisé du site. L'illustration ci-dessous montre le résultat de l'analyse d'une légende d'OpenEdition par l'application Entity Fishing. Le nom de l'auteur (Santiago Ramón Y Cajal) est correctement rattaché à son identifiant sur Wikidata.

EntityFishing est généralement plus performant sur les textes assez normalisés comme les références bibliographiques intégrées dans les légendes.



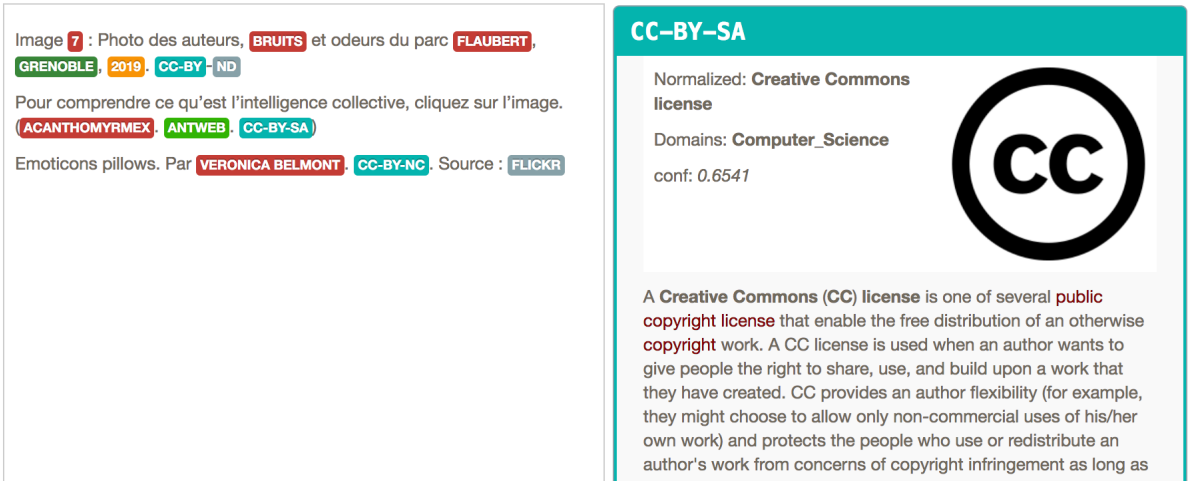

Résultat de l'analyse d'une légende d'OpenEdition par l'application Entity Fishing

Dans l'échantillon annoté, la détection d'entités nommées permet effectivement d'extraire automatiquement des ayants droit. Cependant, rien ne permet de distinguer ces acteurs des autres personnes ou des autres organisations mentionnées dans le texte des légendes. Le tableau ci-dessous met en parallèle les annotations manuelles (colonnes Métadonnées auteur et Métadonnées ayant droit), qui correspondent à la "vérité terrain", et les entités identifiées automatiquement sur quelques images. Les entités correctes apparaissent sur fond vert et représentent manifestement une part minoritaire de l'ensemble des entités repérées.

| Image | Métadonnées auteur | Métadonnées ayant droit | Entité identifiée par EntityFishing |
|---|---|---|---|
| 1.064ms2_image-55.png | | | Joseph d' |
| 1.0ve813_img-3.jpg | Charles Clifford | RAH_Domaine public | Charles Clifford |
| 1.12hgan_img-29.jpg | | | LE MÂTIN |
| 1.1cvpky_img-22.jpg | Thibaut Ruggeri | | Thibaut Ruggeri |
| 1.26w8by_img-1.jpg | Œuvre de J. van der Heyden (de 1605) | | Marie de Médicis |
| 1.26w8by_img-1.jpg | Œuvre de J. van der Heyden (de 1605) | | Jean Ziarnko |
| 1.2k76gv_img-1.jpg | Bret Hartman | | Alex Honnold |
| 1.3fbx9q_img-2.jpg | Yanik Le Guillou | | Y. Le Guillou |

| | | | |
|---|---|---|---|
| 1.3r9g5r_cave-med-600x335.jpg | S. Le Maho | | S. Le Maho |
| 1.3tp2cj_img-4.jpg | Laurène Moraglia | | Bibliothèque Ventilateur |
| 1.3znfpc_img-2.jpg | | Domaine public | Charles Borromée |
| 1.46eopt_img-2.jpg | Jean Montariol | | Jean Monariol |
| 1.4a7g1x_img-5.jpg | | | Sierra de Los Orgaos |
| 1.4a7g1x_img-5.jpg | | | Majaguas-Cantera |
| 1.4a7g1x_img-5.jpg | | | Jean-Bernard |
| 1.5hobzc_img-13.jpg | | | Cliché Bibliothèque |
| 1.7chdr8_Fig-21bis-500x334.jpg | Thomas Sagory | | Thomas Sagory |

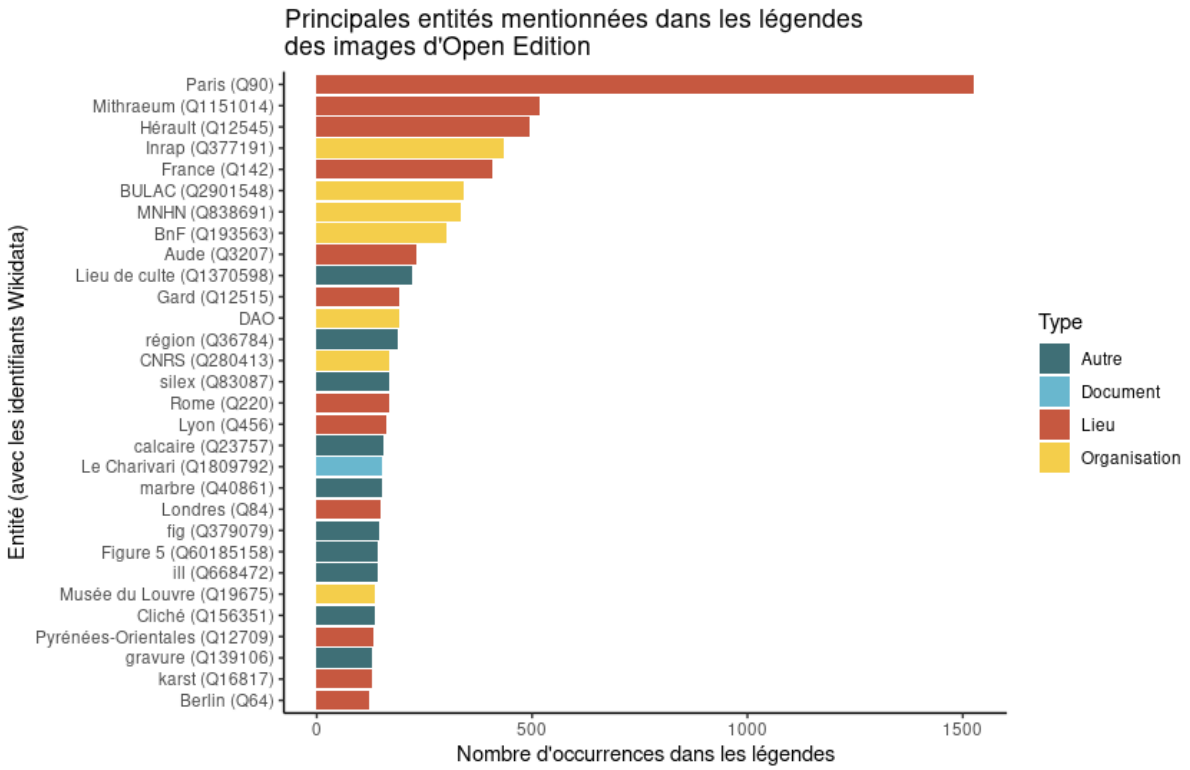
EntityFishing ne permet pas seulement d'identifier des ayants-droits potentiels mais aussi d'autres entités qui qualifient et contextualisent l'attribution. La capture d'écran suivante montre que les licences Creative Commons sont correctement extraites, même si la classification reste perfectible (elles sont catégorisées comme un "concept des sciences informatiques").



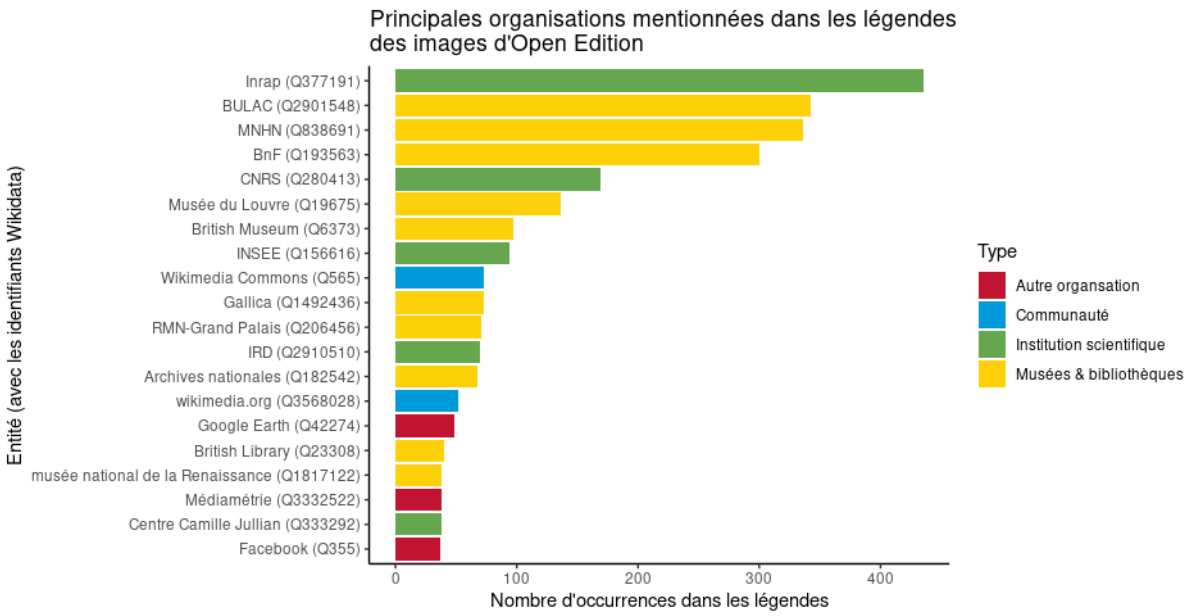

Classification de trois légendes
avec des licences Creative Commons dans EntityFishing

**Nous avons appliqué la reconnaissance d'entités nommées (et leur liaison) à l'ensemble des légendes d'OpenEdition, soit environ 50 000 images mentionnant au moins une légende au sens strict ou un crédit. Le graphique ci-dessous présente les entités nommées les plus fréquemment mentionnées par type.** Les sigles "Q" entre parenthèses correspondent à l'identifiant Wikidata. Bien que toutes ces entités ne soient pas des ayants droit, l'agrégation

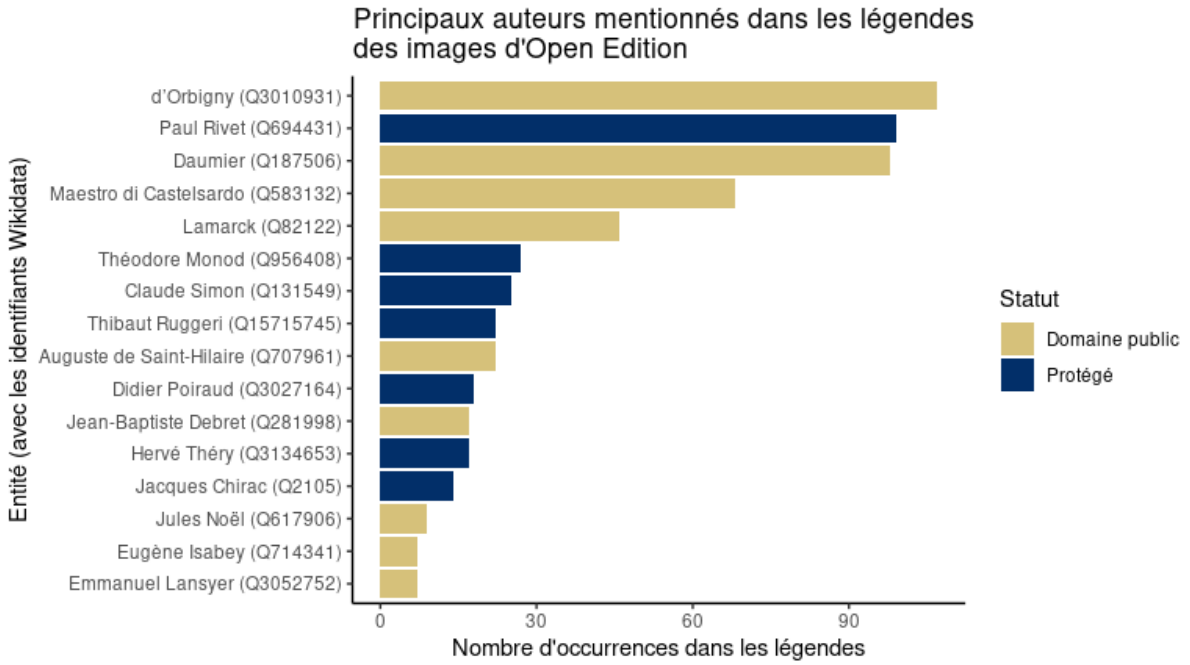
donne un aperçu des institutions et des organisations qui fournissent les images utilisées en SHS.




Dans un second graphique, nous avons effectué une focale rapprochée sur les "organisations". Nous avons introduit une typologie manuelle distinguant les institutions scientifiques, les institutions patrimoniales (musées et bibliothèques), les communautés sous licence libre (Wikimédia Commons) et d'autres organisations. Même si toutes ces organisations ne sont pas forcément citées en tant qu'ayants droit à proprement parler, **ces résultats quantitatifs montrent l'importance prise par les musées, les bibliothèques et les institutions scientifiques dans les pratiques visuelles des sciences humaines et sociales françaises. Par contraste, les banques d'image n'apparaissent pas alors qu'elles jouent un rôle essentiel dans les productions des entreprises ou des médias.**




Le dernier graphe donne la distribution par auteurs. La publication d'études monographiques semblent expliquer la forte présence d'auteurs peu connus (comme le botaniste d'Orbigny ou le Maestro di Castelsardo, un peintre italien anonyme de la Renaissance). Nous avons distingué les auteurs dans le domaine public des auteurs protégés sur la base des dates de décès mentionnés sur Wikidata : **cette typologie montre une prévalence d'auteurs dont les droits patrimoniaux ont expirés.** De plus, deux auteurs encore vivants (Thibaut Ruggeri et Hervé Théry) sont des chercheurs publiants qui peuvent être auteurs ou co-auteurs des productions scientifiques.

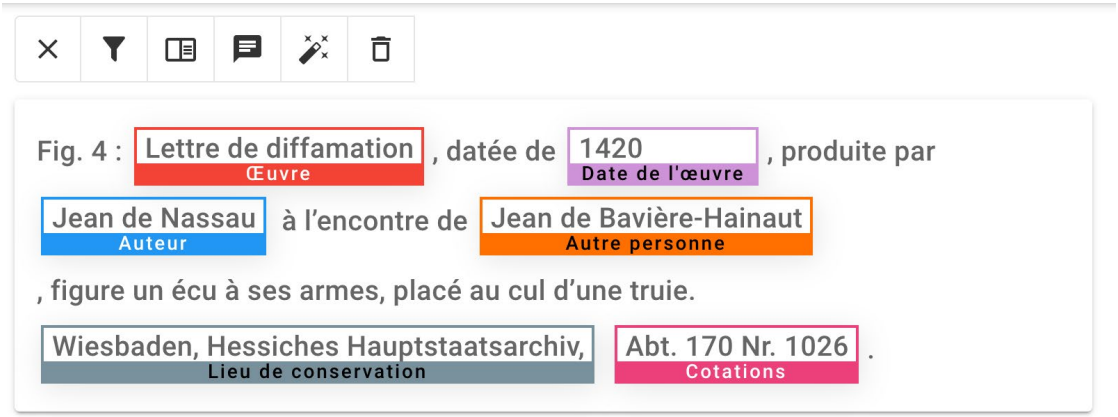



**La répartition des auteurs et des organisations les plus cités dans les légendes d'OpenEdition semble suggérer une prédilection relative pour les productions patrimoniales dans le domaine public, les contenus créés par des collègues ou par les auteurs eux-mêmes affiliés à des**

**institutions scientifiques ou les créations de communautés sous licence libre. Il sera intéressant de voir si cette distribution persiste après la mise en œuvre des licences collectives.**

**Nous avons enfin exploré la possibilité d'entraîner un nouveau modèle de détection d'entité nommée qui intègre une distinction entre les ayants droit et les autres personnes et organisations mentionnées.** De nouveaux programmes d'apprentissage profond (*deep learning*) sont aujourd'hui suffisamment performants pour être mobilisés pour des usages exigeants : Etalab a ainsi récemment développé un outil de pseudonymisation automatisée avec Flair, qui rend possible la diffusion de documents contenant originellement des données personnelles privées[15].

Nous avons préparé un petit corpus d'entraînement de 200 légendes extraites d'OpenEdition avec l'outil d'annotation Doccano en définissant des entités spécifiques : "Auteur" ou "Organisation d'ayant droit" sont ainsi des entités distinctes de "Autre auteur" ou "Autre organisation" ou de "Lieu de conservation". Dans l'illustration ci-dessous, l'auteur de la lettre représentée (Jean de Nassau) est ainsi distingué de la personne visée dans ce document (Jean de Bavière-Hainaut).




**La création d'un modèle avec Flair à partir de ces annotations n'a pas donné de résultats probants : le corpus annoté était finalement trop restreint pour obtenir des données fiables.** Ce travail exploratoire a néanmoins largement contribué à imaginer le workflow automatique que nous proposons dans la dernière partie du rapport.

## Identification des reprises d'images

**Les données internes du modèle de classification permettent également de construire un moteur de recherche d'image inversée *ad hoc* : il est possible de repérer les reprises d'images**

[15] Pour ce test expérimental nous avons repris le protocole décrit dans le manuel de *Pseudonymisation par IA* créé par Etalab : https://guides.etalab.gouv.fr/pseudonymisation/en-pratique/.

**d'un corpus précis dans l'ensemble du corpus.** Cette méthodologie a été initialement expérimentée dans le cadre de travaux historiques sur de grands corpus d'archives patrimoniales : la détection des reprises d'image permet de reconstituer les réseaux de circulation des productions visuelles[16]

Nous avons mené ce test sur un corpus de 1 000 images (de petite taille) de la Réunion des Musées nationaux (RMN) présentes dans les principales galeries de son site web. L'annotation de la phase 2 a montré que ces collections étaient étonnamment peu utilisées, sans doute en raison des restrictions d'usage.

**L'outil a permis de réidentifier 14 reprises d'images (après une vérification manuelle des 100 premiers candidats).** Cette approche fonctionne même lorsque l'image est recadrée ou que la balance des couleurs est différente.
**L'identification des reprises d'images permet de se dispenser des données contextuelles incluses dans la publication scientifique : même en l'absence d'attribution, il est possible de retrouver les ayants droit potentiels en mobilisant les données mises à disposition dans le corpus de référence. Cependant, la méthode est évidemment limitée par la nécessité de définir en amont un corpus de reprises potentielles.**

L'observation précise des cas de reprises met en évidence une grande variété des pratiques d'attribution, qui reflètent généralement la culture disciplinaire du document ou l'expérience de l'auteur (simple étudiant déposant un mémoire de recherche ou bien chercheur confirmé). Trois exemples ci-dessous donnent un bon aperçu de cette diversité :

- Dans le premier cas[17], l'image contient une légende très détaillée attribuant non seulement l'image à l'institution (RMN, Musée du Louvre) mais aussi au photographe à titre individuel : l'autrice du document est une chercheuse en histoire de l'art, spécialiste de l'histoire des collections danoises ;
- Dans le deuxième cas[18], le Louvre est mentionné sans spécifier s'il s'agit juste du lieu de conservation de l'œuvre ou du titulaire des droits de reproduction: l'autrice du document est une étudiante en histoire de l'art.
- Dans le dernier cas, il est simplement mentionné le titre de l'œuvre sans plus de précision : l'auteur est spécialisé en urbanisme et ne mentionne le tableau que pour illustrer une étymologie[19].

[16] Pierre-Carl Langlais, "Stéréotypes Viraux", *DH Nord*, 2020, https://www.meshs.fr/page/viral_stereotypes
[17] https://hal-hprints.archives-ouvertes.fr/hprints-02186083/document (p. 28)
[18] https://dumas.ccsd.cnrs.fr/dumas-02950828/document (p. 58)
[19] https://dumas.ccsd.cnrs.fr/dumas-02176781/document (p. 32)

Les attributions diffèrent mais aussi plus radicalement l'affichage de l'image. Une reproduction du Radeau de la Méduse de Géricault a ainsi été volontairement caviardée en tant que « contenu non libre de droit ». Cependant, ce caviardage est assez approximatif : l'image est juste cachée derrière une autre image blanche et peut toujours être extraite du document PDF.

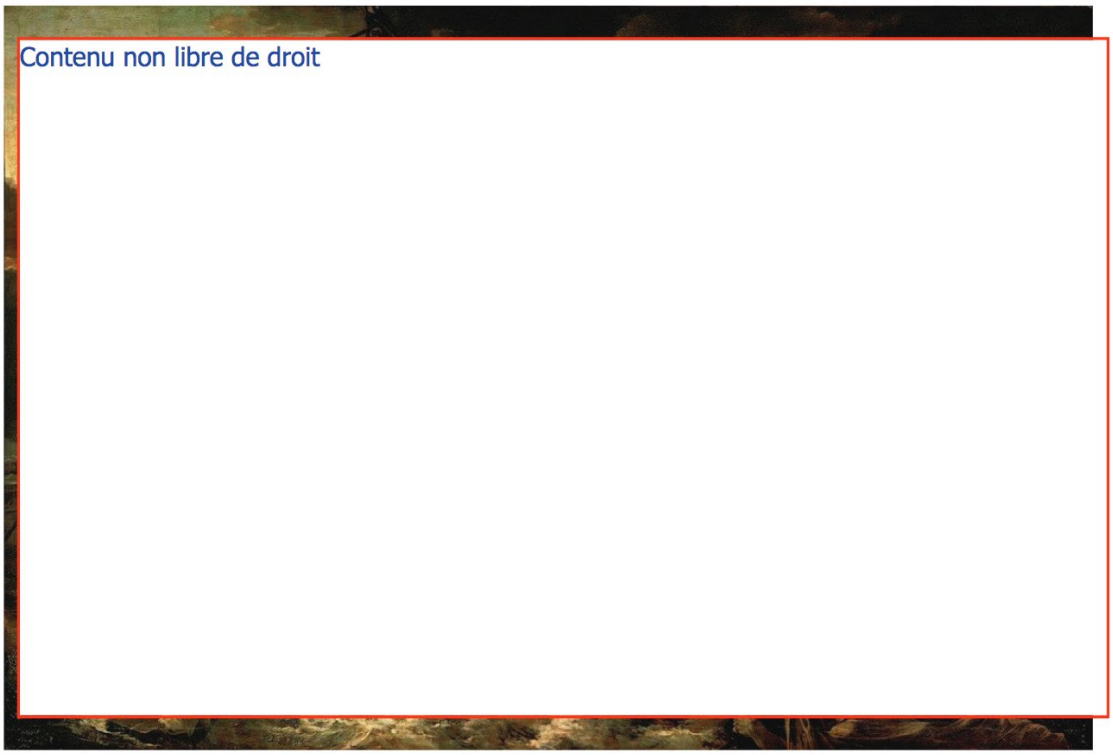



*Figure 9 - Théodore Géricault, Le radeau de la Méduse, 1818-1819.*

Malgré la diversité des résultats, ce test initial a une portée limitée : nous nous sommes limités par commodité aux 1 000 images mises en avant dans les collections en ligne du site de la RMN mais le fonds est considérablement plus large. **Rappelons que l'analyse des entités nommées menées sur les légendes d'OpenEdition a permis d'identifier 136 images mentionnant le Musée du Louvre et 71 images mentionnant la RMN/Grand Palais, soit bien plus que les 14 images recouvrées par l'analyse automatique des reprises.**

## Vers un workflow automatique

**Les tests menés dans cette phase 3 montrent qu'il n'existe pas de méthode unique permettant d'identifier automatiquement le statut des œuvres relevant des arts visuels dans les publications scientifiques.** Chaque méthode décrite ici a ses propres limites :

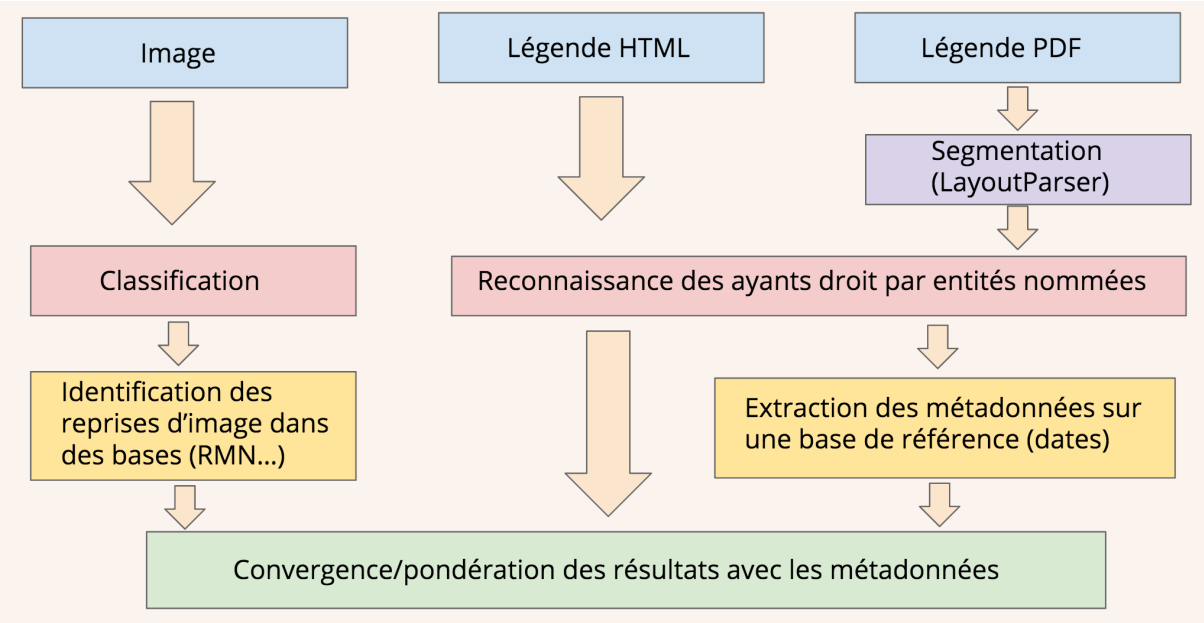
- la classification des images permet d'écarter la plupart des images situées hors du périmètre de l'annotation car ne relevant pas des arts visuels (dans la mesure où il s'agit principalement d'images non figuratives). Les autres cas de figure (domaine public,

création de l'auteur, et image sous droit) sont insuffisamment distingués. À défaut de catégories génériques, il serait cependant envisageable de saisir des ensembles plus circonscrits. Par exemple une version affinée du modèle mis en œuvre dans la phase 1 de l'étude pourrait prévoir une catégorie spécifique pour les gravures anciennes ou les miniatures médiévales (qui relèvent généralement du domaine public) ou pour les schémas d'imagerie médicale (qui ne rentrent généralement pas dans le périmètre des œuvres d'art visuels). Inversement, la classification automatisée laisserait de côté des productions plus incertaines, nécessitant une annotation manuelle, comme les photographies anciennes.

- la classification des légendes donne des résultats très limités, sauf dans une moindre mesure pour les créations des auteurs ;
- l'extraction des entités nommées permet théoriquement d'identifier les auteurs et les organisations titulaires des droits. En pratique, ces données ne sont cependant pas distinguées de l'ensemble des entités ;
- l'identification des reprises d'images fonctionne sans recours aux légendes ou à des métadonnées complémentaires mais est d'emblée limitée à un corpus de comparaison (ici 1 000 images extraits du site de la RMN).

**Les tests montrent aussi qu'il existe des marges d'amélioration substantielles** : nous avons utilisé des modèles et des outils pensés pour d'autres corpus. La segmentation automatique des documents PDF repose sur le modèle *LayoutParser* entraîné sur des articles scientifiques anglo-saxons en sciences et technologies. La détection des entités nommées de EntityFishing est originellement élaborée à partir d'un modèle étiqueté de Wikipédia où la notion d'ayant droit n'a pas vraiment de sens. Et le nombre d'images annotées disponible pour entraîner et tester les algorithmes pourrait être supérieur pour offrir un meilleur résultat.

**Au-delà de ces adaptations, le développement d'un workflow automatique requiert l'utilisation conjointe de plusieurs méthodes.** Le schéma ci-dessous décrit une articulation possible. Le corpus initial est divisé en collection d'images, de documents au format web et de document au format PDF. Chaque collection donne lieu à différents traitements : classification des images, extractions des entités nommées, identification des reprises d'images, croisement avec des métadonnées (discipline, format…), etc.




**La diversification et la redondance des méthodes sera un important facteur pour obtenir des résultats corrects.** L'annotation manuelle a déjà mis en évidence qu'une part substantielle des images n'a pas de légende ou pas de légende exploitables. L'identification de l'image dans un corpus-cible prédéfini ou, à défaut, une imputation probabiliste sur la base de la classification de l'image, du style de la légende ou de la thématique du texte permettront de "rattraper" un certain nombre d'images.

Nous proposons ci-dessous une estimation des tâches nécessaires pour créer ce workflow automatique :

- créer un modèle de segmentation éditoriale des documents PDF (75% du corpus des images figuratives) adapté aux documents SHS publiés en français à partir d'un corpus de page extraites aléatoirement des publications de 2019
- créer un modèle d'extraction d'entités nommées distinguant les ayants droit potentiels des autres personnes et organisations nommées et appliquant d'autres règles permettant de contextualiser l'attribution (licence libre, mention "©")
- affiner le modèle de classification des images avec des catégories plus précises permettant d'exclure plus facilement les images hors critères
- constituer de grands corpus d'images potentiellement reprises (RMN, ADGP)
- créer un méta-modèle probabiliste croisant les données extraites par les étapes précédentes

Nous insistons sur le fait que ce travail ne peut pas être uniquement technique : les modèles doivent être couplés à des règles explicites permettant de déterminer comment il convient de traiter tel ayant droit, telle catégorie d'image, ou tel nom d'auteur.

**Ce travail aurait du sens dans le cadre d'un projet plus général d'amélioration de l'indexation**

**des productions scientifiques françaises en SHS**[20]. La phase 1 de l'étude a déjà permis de constituer un corpus sans précédent des publications en libre accès pendant une année. Une grande partie de ce corpus est au format PDF et n'est pas proprement indexé dans des moteurs de recherche généraliste ou spécialisé : 75% des images figuratives viennent des deux grandes collections en PDF, à savoir le CCSD et les thèses. Au-delà du cas des légendes, la segmentation éditoriale permet ainsi de rendre visible de nombreux éléments importants des documents en PDF : extraction des tables et jeux de données, découpage en titre et en sous-titres, identification des références et des bibliographies. Au-delà de l'identification des ayants droit, l'extraction des entités nommées permet aussi de relier une production scientifique au principal thème traité.

Pour illustrer la valeur ajoutée d'un tel travail, nous proposons dans l'image suivante de simuler une version future du moteur de recherche Isidore qui serait augmentée par les analyses automatisées du corpus. Les métadonnées standard sont enrichies par l'extraction des jeux de données (ici une liste du corpus étudié dans la thèse), l'identification d'images documentées par leur légende (avec des hyperliens vers les personnes et organisation mentionnées lorsqu'elles peuvent être rattachées à une base de référence) et l'indexation d'entités nommées (les personnes et les organisations les plus mentionnées). C'est en tous cas une perspective ouverte par le travail expérimental mené dans la phase 3.




[20] Justement, le Consortium de moyens mutualisés pour des services et données ouvertes en SHS (COMMONS), porté par OpenEdition, Métopes et Huma-Num, a été sélectionné par l'appel à manifestations d'intérêt EquipEx+ du Programme d'investissements d'avenir (PIA 3) pour améliorer l'accès aux publications et aux données, ainsi que la liaison entre publications et données.

# Phase 4 : estimation du nombre d'images dans le champ de mesure pour l'ensemble du corpus

Cette section présente le calcul estimant, à partir des résultats de l'annotation manuelle de la phase 2, le nombre d'images dans le champ de la mesure publiées dans le corpus 2019. La même méthode peut également être itérée à partir d'une nouvelle extraction Isidore : en effet, elle se présente sous la forme d'une série de documents exécutables sous R.

## Encadré : à propos des documents exécutables

Les documents exécutables comprennent une série d'instructions codées dans le langage R, encadré par du texte en langue naturelle qui permet de décrire chaque étape, et vont générer ensemble un rapport directement communicable (au format HTML, DOCX ou PDF). Ils peuvent être téléchargés pour être exécutés dans l'environnement R Studio et modifiés dans le respect de la licence MIT : https://gitlab.huma-num.fr/planglais/images-usages-isidore

Comme l'indique le fichier d'aide (`README.md`) présent au bout du lien ci-dessus, l'estimation peut être produite à partir de trois carnets de code (ou *notebooks*) : `Récupération_du_corpus.Rmd` (qui permet de charger de nouvelles métadonnées d'Isidore), `Application_des_modèles_à_un_nouveau_corpus.Rmd` (qui produit une estimation du nombre d'images à partir d'une modélisation pondérée du nombre de documents) et `Projection_de_la_répartition.Rmd` (qui généralise les résultats de l'analyse manuelle concernant le statut légal des images). Les autres carnets de code documentent la création des modèles et des instruments statistiques utilisés. Il n'est pas nécessaire de les utiliser sauf pour modifier les paramètres des modèles ou pour mieux comprendre les choix méthodologiques effectués en amont.

Chaque carnet de code s'ouvre de préférence dans RStudio. Ils peuvent être utilisés sans connaissance préalable du langage R : il suffit d'exécuter successivement chaque cellule de code en cliquant sur le bouton `Run` et modifier au besoin certaines variables signalées dans le texte.

### Récupération de corpus

Ce carnet récupère les métadonnées pour un corpus d'une année sur Isidore. Il doit être impérativement exécuté en premier. Pour ne pas affecter le bon fonctionnement d'Isidore, la récupération des données est étalée dans le temps. Il est préférable de lancer ce carnet de

code d'une traite (en cliquant sur `Run` > `Run All`) puis de laisser tourner R pendant environ une dizaine d'heures.

### Application des modèles à un nouveau corpus

Ce carnet permet d'estimer le nombre d'images en extrapolant les fréquences statistiques observées dans la présente étude (les "modèles") à un nouveau corpus de métadonnées obtenu depuis Isidore.

Les modèles étant basés sur les disciplines et les plateformes, ils s'adaptent naturellement à l'évolution de la physionomie du corpus Isidore, et gardent leur valeur même si telle plateforme ou telle discipline prend un poids prépondérant.

### Projection de la répartition

Ce carnet permet de généraliser la ventilation statistique des champs annotés manuellement sur l'échantillon des images de l'année 2019. L'échantillon aléatoire a été produit à la phase 2 à partir d'une sélection stratifiée du corpus, visant à en respecter les caractéristiques : à la différence des sondages d'opinion classiques, il n'y a pas eu de biais de sélection. Il est ainsi possible de donner une estimation directement en utilisant la formule classique du calcul de la marge d'erreur sans opérer de correction ou de redressement (comme la méthode des "quotas").

Cette ventilation peut être directement appliquée au nombre d'images estimées par le carnet précédent Application des modèles à un nouveau corpus.Rmd. Cette dernière estimation tient déjà compte de la stratification du corpus (en discipline, en format et par plateforme). En l'absence d'évolution significative des pratiques d'utilisation des images, l'extrapolation des ventilations observées sur l'échantillon de 2019 (considérée comme l'année de référence pour cette étude) devrait rester sensiblement exacte.

Le texte qui suit a été produit par le document exécutable `Projection de la répartition.Rmd`, d'où la présence de code, de texte, et de figures produites par l'exécution du code : c'est ce rapport que vous obtiendrez, avec des valeurs à jour, quand le document exécutable sera exécuté à nouveau (sur le corpus 2019 ou sur toute autre année).

## Préparation des données

Nous ouvrons le fichier d'annotation et enregistrons le nombre total d'images annotées.

```
options(scipen=999)
```

```
library(tidyverse)
annotation = read_tsv("echantillon_annote.tsv")
total_annotation = nrow(annotation)
```

Nous calculons les proportions en pourcentage (ou plus exactement en "pour 1") :

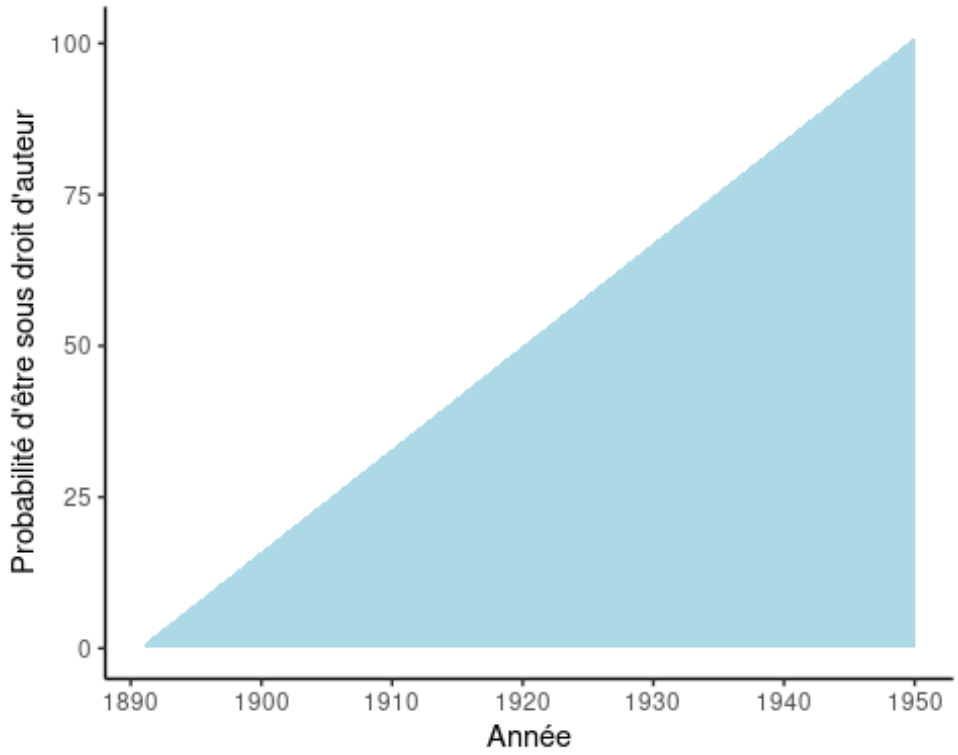
```
sum_annotation = annotation %>%
  count(status_image, name = "image") %>%
  mutate(proportion = (image/sum(image)))
sum_annotation
```

| Statut | Image | Proportion |
|---|---|---|
| Création d'un auteur associé | 69 | 4.53 |
| Création de l'auteur | 296 | 19.45 |
| Hors droit d'auteur | 124 | 8.15 |
| Hors MESRI | 153 | 10.05 |
| Hors périmètre | 438 | 28.78 |
| Image sous droit | 389 | 25.56 |
| Licence libre | 32 | 2.10 |
| Publication à visée lucrative | 21 | 1.38 |

Nous devons calculer à part le corpus restant pondéré en agrégeant les probabilités de droit d'auteur. Ce calcul tient compte des œuvres orphelines qui peuvent être datées entre 1890 et 1950 et suit la formule linéaire :

$$((Année\ de\ publication * 0.017) - 32.14) * 100$$

```
ggplot(tibble(year = 1891:1950) %>% mutate(estimation = ((year*0.017)-
32.14)*100), aes(year, estimation)) + geom_area(fill = "lightblue") +
theme_classic() + labs(x="Année", y="Probabilité d'être sous droit
d'auteur")
```




```
annotation_corpus_pondere = annotation %>%
  filter(status_image == "Image sous droit") %>%
  summarise(image = sum(Probabilite_DA)) %>%
  mutate(proportion = image/total_annotation) %>%
  mutate(status_image = "Image sous droit pondérée")
sum_annotation = sum_annotation %>% bind_rows(annotation_corpus_pondere)
sum_annotation
```

| Statut | Image | Proportion |
|---|---|---|
| Création d'un auteur associé | 69.000 | 4.53 |
| Création de l'auteur | 296.000 | 19.45 |
| Hors droit d'auteur | 124.000 | 8.15 |
| Hors MESRI | 153.000 | 10.05 |
| Hors périmètre | 438.000 | 28.78 |
| Image sous droit | 389.000 | 25.56 |

| | | |
|---|---|---|
| Licence libre | 32.000 | 2.10 |
| Publication à visée lucrative | 21.000 | 1.38 |
| Image sous droit pondérée | 380.693 | 25.01 |

## Calcul de la marge d'erreur et des estimations

Nous calculons la marge d'erreur en suivant la formule classique des sondages sur la base d'un intervalle de confiance à 95% soit :

$$z_{0.95} = 1.96$$

$$n = total\ des\ annotations$$

$$P = proportion\ en\ \%$$

Le calcul à effectuer est donc le suivant :

$$z_{0.95}\sqrt{\frac{\sigma_P^2}{n}} = 1.96\sqrt{\frac{\sigma_P^2}{1522}}$$

```
sum_annotation = sum_annotation %>%
  mutate(margin_error = 1.96*sqrt((proportion*(1-
proportion))/total_annotation)) %>%
  mutate(proportion = round(proportion*100, 2), margin_error =
round(margin_error*100, 2))
```

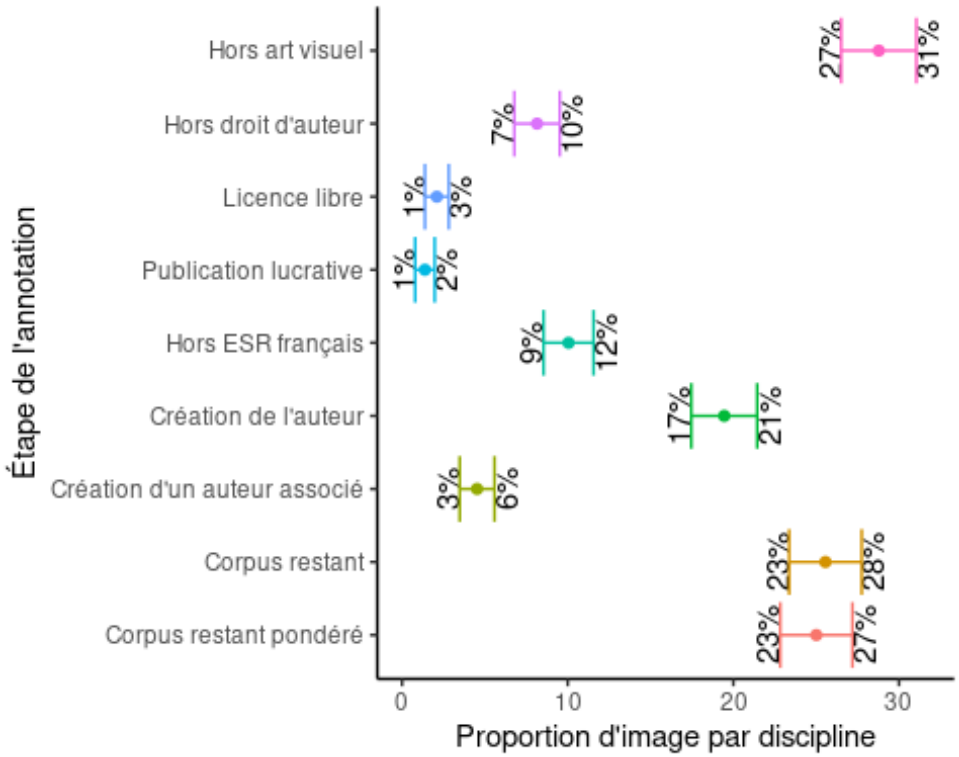
Nous allons maintenant visualiser les résultats. Avant cela nous changeons le nom du "statut" des images pour qu'ils soient plus explicites et définissons un ordre hiérarchique :

```
order_status = tibble(status_image = c("Hors périmètre", "Hors droit
d'auteur", "Licence libre", "Publication à visée lucrative", "Hors MESRI",
"Création de l'auteur", "Création d'un auteur associé", "Image sous droit",
"Image sous droit pondérée"),
                      status_id = 1:9,
                      status_image_label = c("Hors art visuel", "Hors droit
d'auteur", "Licence libre", "Publication lucrative", "Hors ESR français",
"Création de l'auteur", "Création d'un auteur associé", "Corpus restant",
"Corpus restant pondéré"))

sum_annotation = sum_annotation %>%
  inner_join(order_status, by=c("status_image"))
```

La visualisation donne un aperçu de la valeur minimale et maximale estimée de chaque image sur la base de l'intervalle de confiance à 95%. Par exemple pour le corpus restant pondéré la marge d'erreur est de 2.18% : sur la base d'une proportion de 25.01%, l'estimation se trouve dans une fourchette allant de 22.83% à 27.19%.

```
ggplot(sum_annotation, aes(reorder(status_image_label, -status_id),
proportion, color = reorder(status_image_label, -status_id))) +
      geom_text(aes(label = paste0(round(proportion-margin_error), "%"),
y=(proportion-margin_error)-0.7), angle = 90, color = "black") +
    geom_text(aes(label = paste0(round(proportion+margin_error), "%"),
y=(proportion+margin_error)+0.7), angle = 90, color = "black") +
    geom_point() +
    geom_errorbar(aes(ymin = proportion-margin_error,
ymax=proportion+margin_error)) +
    guides(color = FALSE) +
    coord_flip() +
    theme_classic() + labs(x="Étape de l'annotation", y="Proportion d'image
par discipline")
```




Nous enregistrons ces estimations :

```
write_tsv(sum_annotation, "summary_annotation.tsv")
```

## Extrapolation à l'ensemble du corpus.

Pour l'année 2019, nous avons récolté et catégorisé automatiquement l'ensemble des images classées comme des représentations figuratives ou comme des documents. Ces résultats se trouvent dans le fichier `corpus_image_figuratif_document.zip` :

```
corpus_image = read_tsv("corpus_image_figuratif_document.zip")
corpus_image %>% head(100)
```

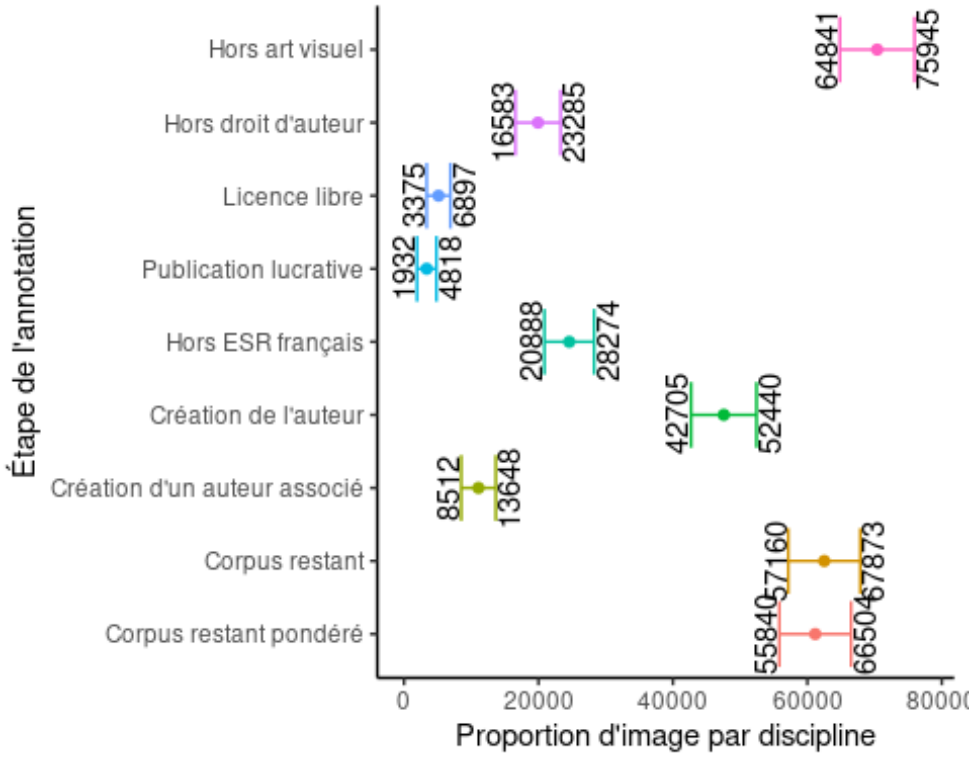
Nous avons au total 244 589 images. Nous pouvons extrapoler directement les estimations et les marges d'erreur calculées sur la base de l'échantillon annoté :

```
total_image_corpus = nrow(corpus_image)
sum_annotation = sum_annotation %>%
  mutate(corpus_image = total_image_corpus*(proportion/100)) %>%
  mutate(corpus_margin = total_image_corpus*(margin_error/100))
sum_annotation %>% select(Statut = status_image_label, Estimation =
corpus_image, "Marge d'erreur" = corpus_margin)
```

| Statut | Estimation | Marge d'erreur |
|---|---|---|
| Hors art visuel | 70,392.71 | 5,552 |
| Hors droit d'auteur | 19,934.00 | 3,351 |
| Licence libre | 5,136.37 | 1,761 |
| Publication lucrative | 3,375.33 | 1,443 |
| Hors ESR français | 24,581.19 | 3,693 |
| Création de l'auteur | 47,572.56 | 4,867 |
| Création d'un auteur associé | 11,079.88 | 2,568 |
| Corpus restant | 62,516.95 | 5,356 |
| Corpus restant pondéré | 61,171.71 | 5,332 |

Nous créons la visualisation pour avoir un aperçu d'ensemble des estimations et des marges d'erreur :

```
sum_annotation %>%
  ggplot(aes(reorder(status_image_label, -status_id), corpus_image, color =
reorder(status_image_label, -status_id))) +
    geom_point() +
    geom_text(aes(label = round(corpus_image-corpus_margin),
y=(corpus_image-corpus_margin)-2000), angle = 90, color = "black") +
    geom_text(aes(label = round(corpus_image+corpus_margin),
y=(corpus_image+corpus_margin)+2000), angle = 90, color = "black") +
    geom_errorbar(aes(ymin = corpus_image-corpus_margin,
ymax=corpus_image+corpus_margin)) +
    guides(color = FALSE) +
    coord_flip() +
    theme_classic() + labs(x="Étape de l'annotation", y="Proportion d'image
par discipline")
```




La visualisation donne de nouveau un aperçu de la valeur minimale et maximale estimée de chaque image sur la base de l'intervalle de confiance à 95%. Ainsi pour le corpus restant la marge d'erreur est de 5 332 images : **sur la base d'une estimation à 61 172 images dans le champ de la mesure, cela représente une fourchette entre 55 840 images et 66 504 images dans le champ de la mesure**.

# Phase 5 : dénombrement des images du portail Persée

## Objectifs

Le portail Persée numérise et met en ligne un vaste corpus (plus de 870 000 documents librement accessibles réparties entre 300 revues ou "collections") du patrimoine des sciences humaines et sociales francophones. Ses collections rétrospectives remontent jusqu'aux années 1870 (exemple des *Annuaires de l'École pratique des hautes études* et des *Archives Parlementaires de la Révolution Française*).

La phase 5 vise à dénombrer les images entrant dans le champ de la mesure :

- parmi les images caviardées qui figurent dans les collections rétrospectives de Persée ;
- parmi les images, caviardées ou non, qui figurent dans les collections 2019 de Persée.

Le dénombrement repose sur un échantillon traité manuellement et extrapolé de manière statistique.

## Construction de l'échantillon

Pour la construction de l'échantillon, nous avons repris pour l'essentiel les règles déjà appliquées lors de la phase 1.

Ainsi, lors de la construction du corpus à la phase 1, étaient exclus les documents dont le texte intégral n'était pas disponible en accès ouvert (puisqu'ils ne sont pas concernés par la mesure). De la même manière, nous avons exclu les documents qui ne sont pas disponibles à la consultation, c'est-à-dire ceux qui sont remplacés par la mention "En raison d'une interdiction de diffusion de la ressource consultée, le contenu de cette page peut être partiellement ou totalement masqué." Ces documents sont signalés avec une icône barrée et bordée de rouge :



Aussi, lors de la construction du corpus à la phase 1, étaient exclues les publications non éditées en France puisqu'elles ne sont pas concernées par la mesure. Or plusieurs revues de Persée sont éditées à l'étranger. Les informations sur l'éditeur disponibles dans les métadonnées fournies nous ont permis d'identifier facilement ces cas :

| Revue | Éditeur | Pays |
|---|---|---|
| *Mélanges de la Casa de Velázquez* | Madrid : Editions de la Casa Velázquez | Espagne |
| *Ebisu - Études Japonaises* | Tokyo : Maison franco-japonaise | Japon |
| *Perspectives Chinoises* | Hong Kong : Centre d'études français sur la Chine contemporaine | Chine |
| *Scriptorium* | Bruxelles : Centre d'Etudes des Manuscrits | Belgique |

Ces quatre revues ne doivent pas être traitées de la même manière. En effet, le Comité de pilotage a identifié trois revues relevant du droit français :

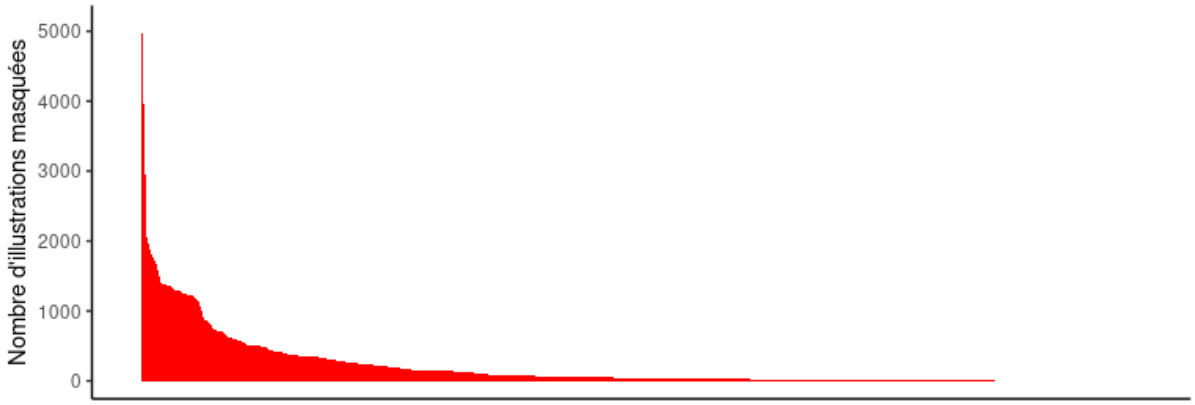
- la Casa de Velázquez est une école française à l'étranger sous tutelle du MESRI
- la Maison franco-japonaise et le Centre d'études français sur la Chine contemporaine sont des instituts français de recherche à l'étranger sous la cotutelle du Ministère de l'Europe et des affaires étrangères et du CNRS.

Par conséquent, au vu de la territorialité de la mesure, le Comité de pilotage convient que seul *Scriptorium* doit être décompté en tant que revue étrangère.

## Échantillonnage des images caviardées des collections rétrospectives de Persée

L'échantillon a été construit en collaboration avec l'équipe de Persée puisque par définition, les images caviardées (appelées également "images masquées") ne sont pas accessibles au public. Selon le nombre d'images caviardées par revue que l'équipe nous a transmis, **45 185 images réparties entre 210 revues sont caviardées sur Persée**.

Comme le montre le graphique ci-dessous, la distribution des images est très inégalitaire : 20 revues concentrent 60% des images caviardées et, inversement, 70 revues comptent moins de 10 images caviardées.




La distribution des images caviardées dans les collections de Persée suit une courbe en "longue traîne".

Néanmoins, si cette statistique agrégée existe, les images caviardées ne sont pas directement identifiables au sein corpus d'images conservé par Persée : elles sont mélangées avec les images effectivement publiées. Pour cibler du mieux possible les images à transmettre par Persée, nous avons dû opérer une sélection préalable des revues contenant un ratio non négligeable d'images caviardées, en appliquant deux règles concomitantes :

- au moins 5% d'images caviardées sur l'ensemble des images de la revue ;
- au moins 2 images caviardées en moyenne par numéro de revue.

Ces seuils sont suffisamment élevés pour cibler les images caviardées (du fait de leur distribution très inégalitaire) et en même temps suffisamment bas pour maintenir un peu de diversité dans la sélection et éviter d'avoir uniquement des articles des mêmes revues ou de la même discipline. Ainsi, 41 revues différentes respectaient ces deux critères (classées ici par nombre d'images caviardées décroissant) :

| **Titre** | **Nombre d'images caviardées** | **Proportion d'images caviardées (%)** |
|---|---|---|
| *Revue de l'Art* | 5113 | 85 |
| *Tiers-Monde* | 2054 | 46 |
| *Revue de géographie alpine* | 1802 | 12 |
| *Population* | 1681 | 6 |
| *Monuments et mémoires de la Fondation Eugène Piot* | 1399 | 14 |
| *Matériaux pour l'histoire de notre temps* | 1379 | 50 |
| *Revue d'histoire de la pharmacie* | 1354 | 19 |
| *Cahiers de civilisation médiévale* | 1283 | 29 |
| *Revue archéologique de Picardie* | 1234 | 5 |
| *Actes de la recherche en sciences sociales* | 1202 | 42 |
| *Revue européenne des migrations internationales* | 1119 | 68 |
| *Revue numismatique* | 842 | 10 |
| *Revue d'économie industrielle* | 749 | 21 |

| | | |
|---|---|---|
| *Communication et langages / Les Cahiers de la publicité* | 705 | 13 |
| *Perspectives chinoises* | 702 | 21 |
| *Revue française d'économie* | 621 | 28 |
| *Archipel* | 609 | 15 |
| *Journal de la Société des océanistes* | 584 | 22 |
| *Scriptorium* | 375 | 9 |
| *Journal de la Société des Américanistes* | 370 | 8 |
| *Livraisons d'histoire de l'architecture* | 361 | 61 |
| *Genèses* | 341 | 64 |
| *Formation Emploi* | 277 | 11 |
| *Espace, populations, sociétés* | 269 | 5 |
| *Revue du monde musulman et de la Méditerranée* | 258 | 17 |
| *Bulletin de la Société Nationale des Antiquaires de France* | 240 | 6 |
| *Mélanges de la Casa de Velázquez* | 236 | 6 |
| *Cahiers du Centre Gustave Glotz* | 221 | 27 |
| *Cahiers du Centre d'Etudes Chypriotes* | 195 | 7 |
| *Annuaire des collectivités locales* | 189 | 15 |
| *Flux* | 164 | 21 |
| *Genesis (Manuscrits-Recherche-Invention)* | 154 | 8 |
| *Sociétés contemporaines* | 149 | 24 |
| *Ebisu* | 139 | 16 |
| *Actes des congrès de la Société des historiens médiévistes de l'enseignement supérieur public* | 133 | 16 |
| *Médiévales* | 127 | 26 |
| *Mètis. Anthropologie des mondes grecs anciens* | 123 | 54 |
| *Recherches sur Diderot et sur l'Encyclopédie* | 81 | 23 |
| *Critique internationale* | 69 | 62 |
| *Gaia : revue interdisciplinaire sur la Grèce Archaïque* | 65 | 24 |
| *Seizième Siècle* | 40 | 17 |

Liste des revues du corpus étudié avec pour chacune le nombre absolu d'images caviardées et leur proportion relativement à l'ensemble des images de la revue.

Comme le montre le tableau ci-dessus, les revues en histoire de l'art sont particulièrement présentes (jusqu'à 85% d'images caviardées pour la *Revue de l'Art*). Au-delà de cette prédominance, le corpus couvre un spectre assez large des publications en sciences humaines et sociales : géographie, histoire, sociologie, littérature, sciences de l'éducation.

Dans un second temps, nous avons tiré au hasard cinq numéros pour chacune de ces revues et avons demandé à Persée les images (caviardées et non caviardées) publiées dans ces numéros, soit 12 822 images.

Nous avons recoupé ce premier tirage avec le texte intégral des articles pour ne garder que les images présentes dans une page contenant la mention "Illustration non autorisée à la diffusion". Cette sélection n'est pas parfaite mais est généralement opérante : dans certains cas, une page peut contenir des images caviardées et des images non caviardées. Nous avons retiré manuellement ces cas de faux-positifs.

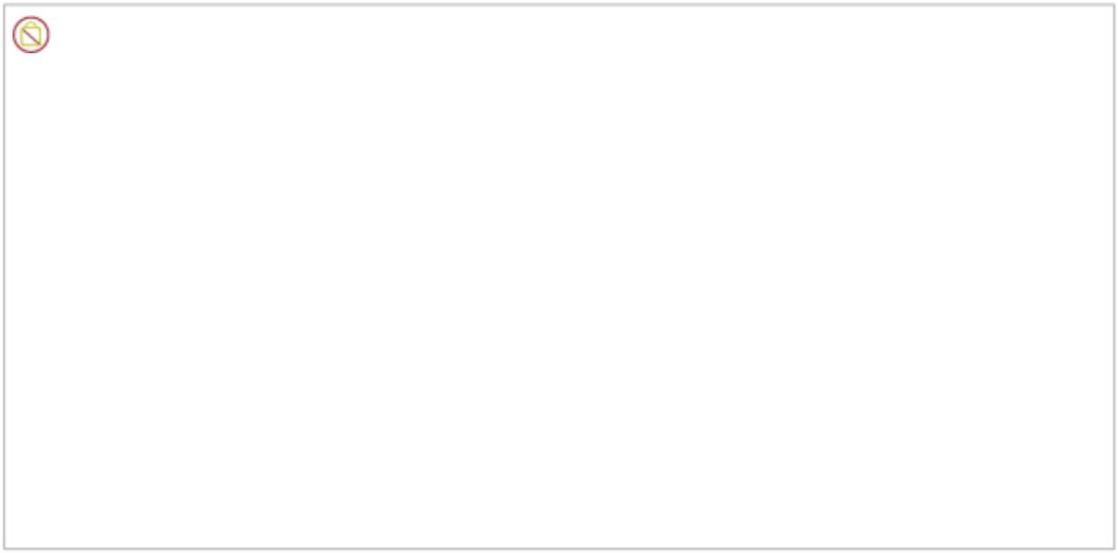
Nous avons ensuite opéré une nouvelle sélection aléatoire de 1 022 images qui ont fait l'objet d'une classification automatisée avec le modèle décrit lors de la phase n°1 de l'étude. Seules les images classées comme des documents (au nombre de 143) et des photographies et peintures (représentations figuratives au nombre de 596) ont été retenues :

| **Classification** | **Nombre d'image** |
|---|---|
| Tableaux, textes, logos & artefacts | 125 |
| Graphe, cartes & schémas | 168 |
| Documents | 143 |
| Photographies & peintures | 596 |

La première catégorie s'explique par le fait que le caviardage de Persée porte aussi sur des objets non visuels comme des tableaux. Par exemple, ce tableau de la féminisation des professions juridiques[21] a été retirée de la version mise en ligne :

[21] Le Feuvre Nicky, Walters Patricia. "Égales en droit ? La féminisation des professions juridiques en France et en Grande-Bretagne". In: *Sociétés contemporaines* N°16, Décembre 1993. pp. 41-62. https://doi.org/10.3406/socco.1993.1140

TABLEAU 2 - LA FÉMINISATION DES PROFESSIONS JURIDIQUES
EN FRANCE ET EN GRANDE-BRETAGNE EN 1989



Au terme du traitement, nous obtenons un **tirage final de 143 + 596 soit 739 images dont 514 ont pu être annotées** (l'annotation a été particulièrement efficace dans le temps imparti en raison de l'exhaustivité des métadonnées fournies par Persée et de la récurrence des sujets de publications, en particulier dans les revues en histoire de l'art).

### Échantillonnage des images des publications 2019 de Persée

Persée a transmis **3 659 images extraites des 487 articles publiés pendant l'année 2019** et référencés par Isidore.

Nous avons opéré ici le même traitement que pour les images du corpus principal constitué lors de la phase 1 de l'étude, dont ce corpus n'est qu'un prolongement. Les images ont été automatiquement catégorisées à partir du modèle de classification, dont **1 665 images ont été classées comme des documents/photographies/peintures**. Quantitativement, cela représente 0,67% du corpus de 248 564 images utilisé pour base de l'annotation à la phase 2 ; par conséquent, afin de garder la même proportionnalité dans l'échantillon annoté manuellement, nous tirons aléatoirement 0,67 % de ces 1 665 images soit 11 images.

## Méthode d'analyse

Nous avons repris exactement les mêmes champs descriptifs, critères d'appréciation et sources de référence que pour la phase 2, à une exception près : les documents étant plus anciens, il s'est avéré plus difficile d'identifier si, à la date de la première publication, au moins un des auteurs était affilié ou associé à un organisme ou établissement d'enseignement supérieur et de recherche public français. ScanR s'avère inopérant car basé sur des données récentes. Dans ce cas nous avons utilisé les notices biographiques du site Persée, elles-mêmes basées sur data.bnf.fr, qui donnent une information (même générale) sur la carrière des

auteurs.

Comme pour la phase 2, nous avons considéré comme organisme ou établissement d'enseignement supérieur et de recherche public français une structure absente du RNSR (Répertoire national des structures de recherche) :

- le Centre de recherches sur les monuments historiques, rattaché à la Médiathèque de l'architecture et du patrimoine (créée par l'Arrêté du 4 janvier 2000 érigeant la médiathèque de l'architecture et du patrimoine en service à compétence nationale).

Considérant, en accord avec le Comité de pilotage de l'étude, que la diffusion sur Persée des revues est une diffusion à but non lucratif, nous avons systématiquement validé le critère "l'image est présente dans une publication à but non lucratif".

## Résultats de l'analyse

### Résultat général

Pour rappel, les principaux champs d'annotations correspondent à un arbre de décision : chaque image passe une série de "tests" successifs. Nous évaluons consécutivement si :

- une image relève des arts visuels
- elle est susceptible d'être protégée par le droit d'auteur
- elle est sous licence libre
- elle est présente dans une publication à but non lucratif
- elle est présente dans une publication d'un auteur affilié à un établissement de l'ESR français
- elle n'est pas une création personnelle de l'auteur
- elle n'est pas une création d'un auteur appartenant au projet ou à l'équipe à l'origine du document.

Le tableau ci-dessous décrit chaque test successif en mentionnant les images conservées ("oui"), les images écartées ("non") et la proportion d'images conservées.

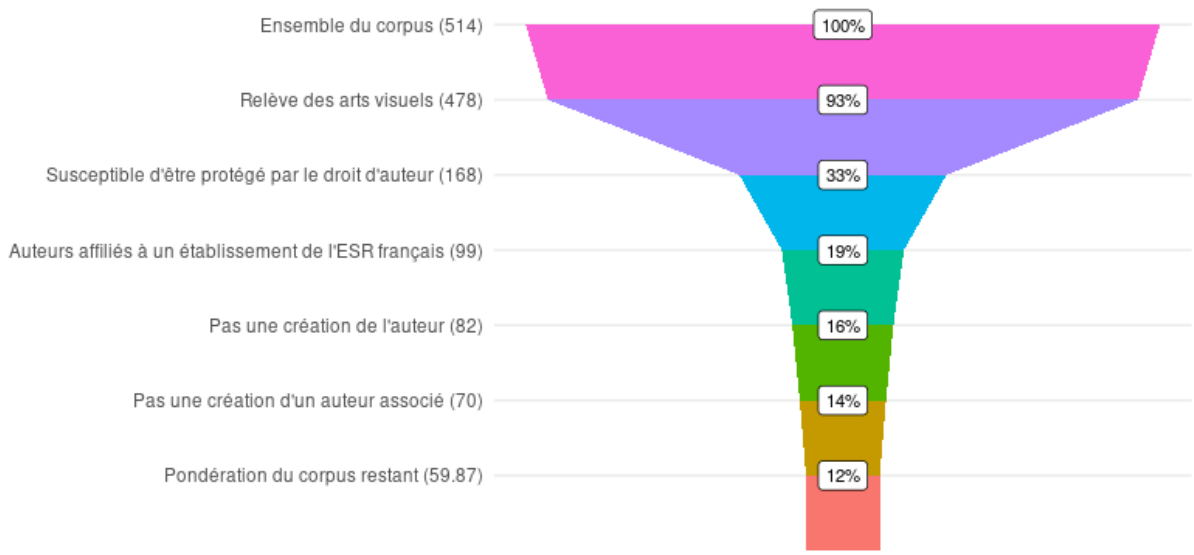
| Critère | Non | Oui | Proportion d'images conservées |
|---|---|---|---|
| Images dans l'échantillon | | 514 | 100 % |
| Relève des arts visuels | 36 | 478 | 93 % |
| Susceptible d'être protégé par le droit d'auteur | 310 | 168 | 33 % |
| N'est pas couvert par une licence libre | 0 | 168 | 33 % |

| | | | |
|---|---|---|---|
| Publication à but non lucratif | 0 | 168 | 33 % |
| Auteurs affiliés à un établissement de l'ESR français | 69 | 99 | 19 % |
| Pas une création de l'auteur | 17 | 82 | 16 % |
| Pas une création d'un auteur appartenant au projet ou à l'équipe | 12 | 70 | 14 % |
| | | | |
| **Sous-total** | | **70** | **14 %** |
| **Sous-total pondéré par la probabilité d'être sous droit d'auteur** | | **59,9** | **12 %** |

Les deux sous-totaux reprennent d'une part le nombre total d'images ayant passé tous les tests, et d'autre part une pondération par la probabilité d'être sous droit d'auteur ; en effet, pour les images datées dont l'auteur n'a pas pu être déterminé, nous avons appliqué une règle de calcul de probabilité telle que définie pour le champ Probabilite_DA (chaque probabilité étant comprise entre 0 et 1, le sous-total pondéré est logiquement inférieur).

Le corpus rétroactif de Persée se caractérise par une **forte proportion d'images non susceptibles d'être protégées par le droit d'auteur**, selon la définition fixée par le Comité de pilotage de l'étude : des images dont les droits patrimoniaux ont expiré et qui ne font mention d'autres droits associés (droit du photographe, droit de l'institution…). Lors de l'application de ce critère, la proportion d'images conservées passe drastiquement de 93% à 31% du corpus.

Cette faible proportion apparaît nettement dans la visualisation ci-dessous, sous la forme d'un entonnoir, montrant l'attrition progressive de l'échantillon après application de chaque critère :




Si seulement un tiers des caviardages porte sur des images susceptibles d'être protégées par le droit d'auteur, cela peut s'expliquer par une combinaison de facteurs propres à la nature du

corpus Persée :

- la prévalence des revues d'histoire de l'art et d'histoire, qui sont fortement utilisatrices d'œuvres visuelles anciennes donc plus susceptibles d'appartenir au domaine public ;
- la faible présence d'indices attribuant aux musées et institutions culturelles dépositaires d'œuvres appartenant au domaine public un droit d'auteur sur la reproduction de ces oeuvres (voir ci-dessous quelques exemples d'attribution au Musée du Louvre, Metropolitan Museum of New-York, Musée national d'Athènes, Musée National des châteaux de Malmaison et Bois-Préau). Dès lors, s'agissant de reproductions à l'identique d'œuvres appartenant au domaine public, elles sont considérées dans l'annotation comme non susceptibles d'être protégées par le droit d'auteur ;
- la politique prudente de Persée en matière de caviardage, comme nous l'a confié sa directrice et comme le prouve le fait que même des tableaux de chiffres soient caviardés.

<table>
<tr>
<td>

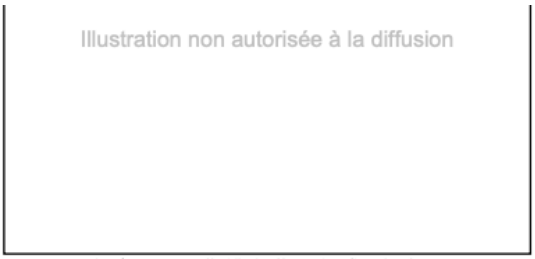
Fig. 1. — La déesse Hathor tendant le collier « menat » au roi Sethôsis Ier, calcaire peint, XIXe dynastie (Musée du Louvre).</td>
<td>

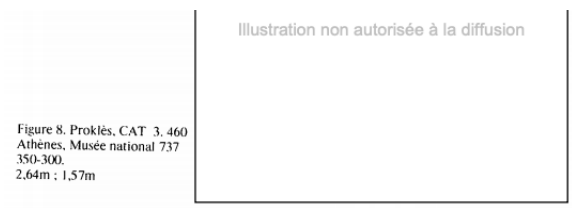
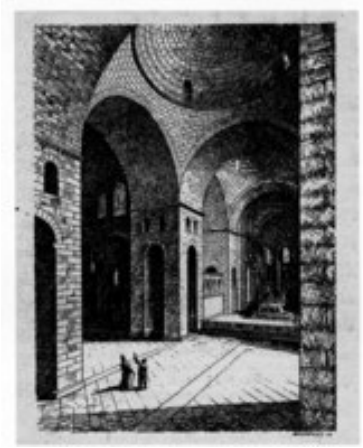
Fig. 3. — Coupe par Hégésibouloss (Musée métropolitain, New-York).</td>
</tr>
<tr>
<td>Boreux Charles. La statue du «serviteur royal » Nofirronpit (Musée du Louvre). In: Monuments et mémoires de la Fondation Eugène Piot, tome 33, fascicule 1-2, 1933. pp. 11-26. https://www.persee.fr/doc/piot_1148-6023_1933_num_33_1_1900</td>
<td>Philippart Hubert. Deux coupes attiques à fond blanc du Musée du Cinquantenaire à Bruxelles. In: Monuments et mémoires de la Fondation Eugène Piot, tome 29, fascicule 1, 1927. pp. 99-136. https://www.persee.fr/doc/piot_1148-6023_1927_num_29_1_1864</td>
</tr>
<tr>
<td>

Figure 8. Proklès, CAT 3, 460
Athènes, Musée national 737
350-300.
2,64m ; 1,57m</td>
<td>

Napoléon à Sainte-Hélène, gravure, Delaistre (Louis Jean Désiré ? 1800-v.1860) (peut-être de la famille d'un statuaire 1746-1822 qui a laissé entre autres pièces un Joseph Bonaparte et des bas-reliefs pour le Panthéon et la colonne Vendôme), s.d., Musée National des châteaux de Malmaison et Bois-Préau.</td>
</tr>
</table>

| | |
|---|---|
| Hoffmann Geneviève. L'expression du temps sur les stèles funéraires attiques. In: *Mètis. Anthropologie des mondes grecs anciens*, vol. 12, 1997. pp. 19-43. https://www.persee.fr/doc/metis_1105-2201_1997_num_12_1_1060 | Rolland Denis, Capdevila Luc. Introduction : France et Belgique, terres d'exil ?. In: *Matériaux pour l'histoire de notre temps*, n°67, 2002. Pour une histoire de l'Exil français et belge, sous la direction de Robert Frank. pp. 1-10. https://www.persee.fr/doc/mat_0769-3206_2002_num_67_1_402378 |

Il convient de noter que dans leur grande majorité, les représentations iconographiques d'œuvres visuelles caviardées sont en noir et blanc dans une qualité assez médiocre (voir tableau ci-dessous) :

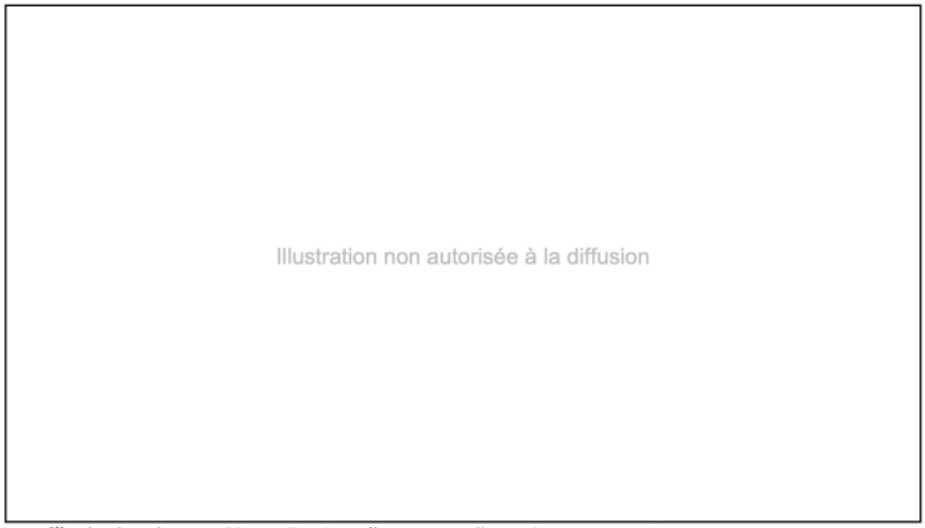
| | |
|---|---|
| <br>Soriguerola. Retable. Détail. La pesée des âmes (Musée de Barcelone) | <br>Saint-Front de Périgueux, vue perspective de la nef publiée dans l'Histoire de l'architecture romane, op. cit., dessin de Corroyer, gravure de Michelet. Cl. Thierry Dechezleprêtre. |
| Durliat Marcel. La peinture romane en Roussillon et en Cerdagne. In: *Cahiers de civilisation médiévale*, 4e année (n°13), Janvier-mars 1961. pp. 1-14. https://www.persee.fr/doc/ccmed_0007-9731_1961_num_4_13_1174 | Gloc Marie. Édouard-Jules Corroyer (1835-1904) : la construction romane, moment décisif dans l'histoire de l'architecture médiévale. In: *Livraisons d'histoire de l'architecture*, n°9, 1er semestre 2005. pp. 99-111. https://www.persee.fr/doc/lha_1627-4970_2005_num_9_1_999 |

<table>
<tr>
<td><br>Photographie de la façade ouest d'Albert Gate House, siège de l'ambassade de France à Londres, MAE, fonds iconographique, Fa Londres.</td>
<td><br>Le retour des cendres de l'Aiglon, L'Illustration, 21 décembre 1940 (couverture)</td>
</tr>
<tr>
<td>Dasque Isabelle. Les hôtels diplomatiques : un instrument de prestige pour la République à l'étranger (1871-1914). In: Livraisons d'histoire de l'architecture, n°4, 2e semestre 2002. pp. 43-68. https://www.persee.fr/doc/lha_1627-4970_2002_num_4_1_914</td>
<td>Rolland Denis, Capdevila Luc. Introduction : France et Belgique, terres d'exil ?. In: Matériaux pour l'histoire de notre temps, n°67, 2002. Pour une histoire de l'Exil français et belge, sous la direction de Robert Frank. pp. 1-10. https://www.persee.fr/doc/mat_0769-3206_2002_num_67_1_402378</td>
</tr>
<tr>
<td><br>Les élèves et le corps enseignant du collège dominicain Captier de Saint- Sébastien (coll Delaunay). Le collège a été fondé en 1903 par les dominicains de Sorèze dans le Tarn (16 religieux français en 1911).</td>
<td></td>
</tr>
<tr>
<td>Delaunay Jean-Marc. L'Espagne, une terre d'accueil pour les Français de l'exil (fin XVIIIe-début XXe s.). In: Matériaux pour l'histoire de notre temps, n°67, 2002. Pour une histoire de l'Exil français et belge, sous la direction de Robert Frank. pp. 36-40. https://www.persee.fr/doc/mat_0769-</td>
<td></td>
</tr>
</table>

| | |
|---|---|
| 3206_2002_num_67_1_402384 | |

## Résultats complémentaires

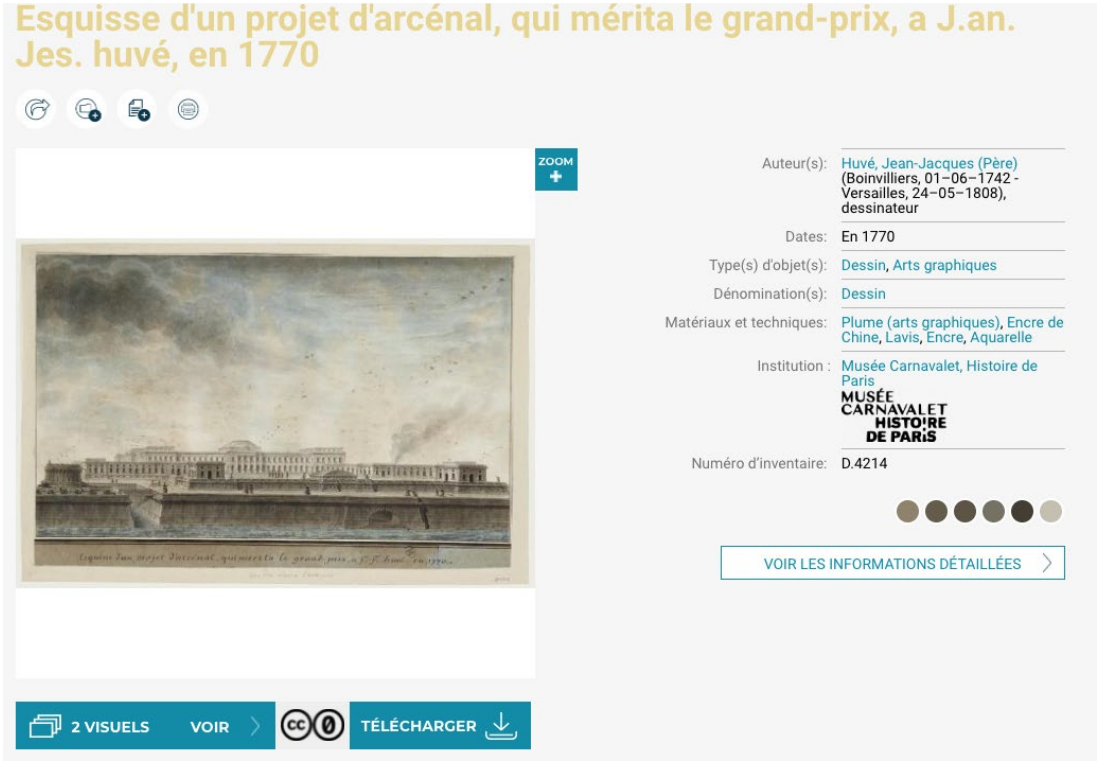
Les images caviardées sont parfois disponibles pour une libre réutilisation auprès de leurs ayants droit. C'est le cas par exemple de cette œuvre dans le domaine public dont le cliché a été pris par Daniel Lifermann pour le compte de la Photothèque des Musées de la Ville de Paris (PMVP) :




Ill. 4 : Jean-Jacques Huvé, Esquisse d'un projet d'arsenal qui mérita le grand prix en 1770, 0,362 × 0,24 m, plume et aquarelle, musée Carnavalet. Cl. PMVP / Lifermann.

La même image est désormais directement disponible sous licence CC0 sur le site parismuseescollections.paris.fr, suite à une démarche récente d' "open content"[22] :

[22] "Open content : plus de 150 000 œuvres des collections des musées de la Ville de Paris en libre accès", 8 janvier 2020, https://www.parismusees.paris.fr/fr/actualite/open-content-plus-de-150-000-oeuvres-des-collections-des-musees-de-la-ville-de-paris-en

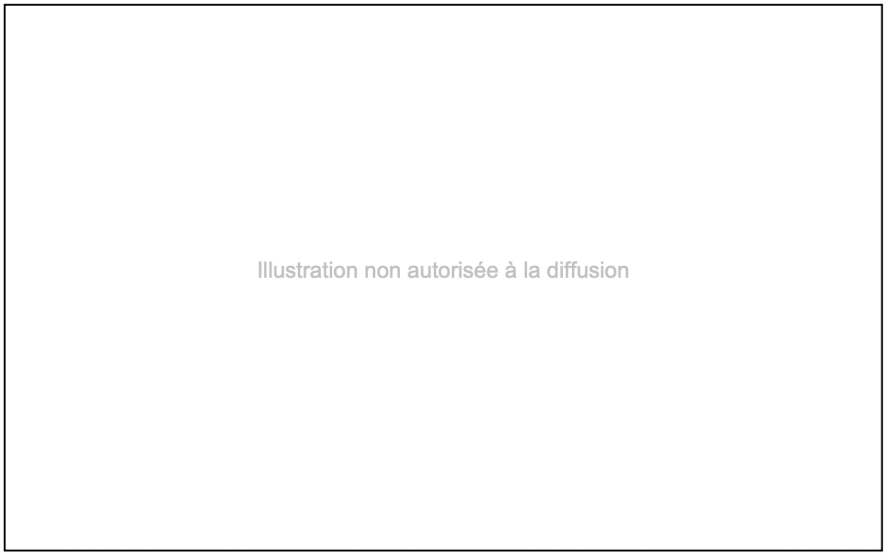



À ce sujet, aucune image sous licence libre n'a été constatée dans l'échantillon.

Aussi, c'est la première fois que nous observons des images du domaine public avec attribution secondaire à l'auteur du document (voir dans le tableau ci-dessous les mentions "Cl. E. Castaner Munoz" et "Cl. Béatrice Bouvier" qui sont des exemples tirés de la même revue *Livraisons d'histoire de l'architecture*). Selon la règle d'annotation définie avec le Comité de pilotage, ces images ont été considérées comme susceptibles d'être protégées par le droit d'auteur et comme des créations de l'auteur.

| |
|---|
| Illustration non autorisée à la diffusion<br>Ill. 5 : Charles Trénet, perspective de l'immeuble et du boulevard, s.d., 25 × 20 cm, mine de plomb, Arch. mun. de Perpignan, 5 S 1/1. Cl. E. Castaner Munoz. |
| Esteban Castaner Munoz, "La « Maison de l'américaine » de l'architecte Claudius Trénet : esthétique |

urbaine et débat stylistique dans l'architecture du début du XXe siècle à Perpignan", *Livraisons d'histoire de l'architecture*, n°7, 1er semestre 2004, pp. 87-98, https://www.persee.fr/doc/lha_1627-4970_2005_num_9_1_995




Ill. 3 : « Autre cité lacustre dans la Nouvelle-Guinée (village de Sowek) », Charles Garnier, *L'Habitation humaine*, Paris, Hachette, 1892, p. 59. Cl. Béatrice Bouvier.

Béatrice Bouvier, "Charles Garnier (1825-1898) architecte historien de L'Habitation humaine", *Livraisons d'histoire de l'architecture*, n°9, 1er semestre 2005, pp. 43-51, https://www.persee.fr/doc/lha_1627-4970_2005_num_9_1_995

Au total, nous dénombrons **36 attributions multiples pour 514 images**, soit une proportion nettement supérieure à la phase 2 qui comptait 15 attributions multiples pour 1 514 images. La majorité relèvent des *Livraisons d'histoire de l'architecture* soit la situation que nous avons décrite ci-dessus. Les autres relèvent essentiellement de photographes d'institutions patrimoniales ou de l'Inventaire général. Un seul cas relève de la représentation d'un bâtiment construit par un architecte, avec double attribution à l'architecte et au photographe.

Nous constatons enfin de façon générale que les métadonnées sur les images que possède Persée sont très précises et complètes, souvent plus que la légende affichée avec l'image. Dans de nombreux cas, ces données peuvent être directement utilisées pour évaluer le statut des images, y compris de manière semi-automatique.

## Dénombrement par extrapolation

En partant sur une extrapolation simple de l'annotation manuelle, **le nombre total pondéré d'images de Persée dans le champ de la mesure serait de 3 904**. L'estimation repose successivement sur le calcul du nombre d'images classées comme des documents ou des images figuratives (72%) et du nombre d'images pondérées dans le champ de la mesure (12%

de ces 72%, soit 8,64% du total) :

| Étape du calcul | Proportion à appliquer | Nombres d'images concernées |
|---|---|---|
| Nombre total d'images caviardées | 100 % | 45 185 |
| Images classées comme des documents ou images figuratives | 72 % | 32 533 |
| Pondération des images de l'échantillon dans le champ de la mesure | 12 % | 3 904 |

En conclusion, **la mise en œuvre de la licence collective étendue selon les termes de l'article 28 de la loi de programmation de la recherche ne permettrait de "libérer" que 4 000 images sur les plus de 45 000 images caviardées figurant dans les collections rétrospectives de Persée**. L'analyse de l'échantillon a fait apparaître en effet que **la protection par le droit d'auteur n'est pas le principal motif de caviardage des images**, bon nombre des œuvres reproduites sur ces images appartenant en réalité au domaine public. Ce sont davantage **les politiques pratiquées par les institutions culturelles et patrimoniales** pour la reproduction des œuvres appartenant à leurs collections (demandes d'autorisation et tarification), ainsi que le principe de prudence appliqué par Persée, qui ont conduit à renoncer à la publication de ces images.
Une politique de "libération" des images caviardées dans les collections rétrospectives de Persée ne saurait donc être opérée de manière systématique, mais devrait s'appuyer sur un travail d'analyse permettant d'identifier:

- les images caviardées qui sont en réalité des tableaux, logos, textes, schémas, qui pourraient être facilement identifiées grâce à la méthode de classification automatisée mise en œuvre dans le cadre de cette étude;
- les images représentant des œuvres qui entreront effectivement dans le champ de la licence collective étendue;
- les images qui n'entrent pas dans le champ de la licence collective étendue, mais qui peuvent être libérées parce qu'elles représentent des œuvres qui appartiennent désormais au domaine public, ou parce que leur auteur est l'auteur de l'article ou un auteur associé.

Pour les images initialement soumises à autorisation et tarification de la part des institutions culturelles et patrimoniales qui les ont communiquées, une mise à jour des conditions désormais pratiquées par ces institutions permettrait de libérer certaines d'entre elles, comme le montre l'exemple de l'image issue des collections des musées de la Ville de Paris.