UNIVERSITÉ DE GENÈVE

Comprendre le Numérique : Cours Transversal
Professeur Yaniv Benhamou

Semestre de printemps
2020-2021

RAPPORT STATISTIQUE

VISUAL CONTAGIONS

GT 6

Sofiane Alloun, Maria Barchenko, Pavel Baruzdin, Céline Belina, Ivana Galic, Angelica Klaus, Rebeka Mali, Nastassia Merlino, Ana Cecilia Reques Araujo, Leah Teague

Référente – Professeure Béatrice Joyeux-Prunel

Assistante – Mme Anna Scius-Bertrand

Genève, juin 2021

Table des matières

Introduction

Le projet *Visual Contagions*[1] a pour ambition d'étudier et de visualiser la circulation mondiale des images grâce à un large corpus d'imprimés, de revues artistiques ou encore des journaux de propagande – et ce, entre 1890 et le début de l'ère d'Internet. Pensées comme un véritable témoin d'un mouvement à l'échelle globale, les contagions visuelles servent de pont entre différentes iconographies rendant ainsi perceptibles des *visual blockbusters*. Le projet a comme vocation de décrire, d'analyser et de rendre visible la circulation des images de manière quantitative en retraçant donc les reproductions, copies ou autres pastiches. L'ambition première est donc d'envisager et de cerner ce qui a contribué à la diffusion d'une image tout en questionnant la circulation face à la mondialisation des cultures et la domination symbolique de certains pays et cultures sur d'autres selon les époques. *Visual Contagions* propose subséquemment d'écrire une histoire inédite de la mondialisation par l'image[2] ainsi que par une approche quantitative et iconographique et s'inscrit dans les Humanités Numériques[3].

Ce rapport statistique s'inscrit dans la continuité du projet mère *Visual Contagions* et permet à un groupe interdisciplinaire de dix élèves de l'Université de Genève de développer un livrable autour de ce sujet le temps d'un semestre. Pour ce faire, une approche duale a été définie permettant de concert d'offrir une analyse quantitative détaillée ainsi qu'un volet plus accessible et pédagogique par le biais d'une carte interactive intégrée au sein d'un site web[4]. À travers l'étude d'un large corpus, l'objectif était de se confronter à une plateforme nouvelle, Explore, qui ouvre la voie à une nouvelle approche des sciences sociales en les confrontant à une analyse quantitative de la mondialisation basée sur des algorithmes, l'analyse et la clusterisation d'images. Ces dernières sont de fait un véritable défi pour l'histoire de la globalisation[5].

Concernant cette première partie du livrable, quelques deux mille revues digitalisées en format IIIF[6] ont amené à se questionner sur nombre de paramètres. Permettant de visualiser et de quantifier les images et leur circulation. Ainsi existe-t-il des routes de la mondialisation ? Sont-elles hétérogènes ? Dépeignent-elles les grands centres artistiques – Paris puis New York ? Est-ce que les flux dépendent

[1] Plus amples informations se trouvent sur le site mère du projet : https://www.unige.ch/visualcontagions/

[2] MITCHELL W J T, *What do Pictures Want? : The Lives and Loves of Images,* Chicago, University of Chicago Press, 2005.

[3] MOUNIER Pierre, « Manifeste des Digital Humanities », Journal des anthropologues [En ligne], 122-123 | 2010, mis en ligne le 01 décembre 2012, consulté le 10 mai 2021. URL : http://journals.openedition.org/jda/3652 ; DOI : https://doi.org/10.4000/jda.3652

[4] Le site web proposé dans le cadre du livrable est accessible à l'adresse suivante : https://unige-cn.wixsite.com/visualcontagions

[5] SUBRAHMANYAM Sanjay, « Historizing the Global, or Labouring for Invention? », *History Workshop Journal,* Vol. 64 (1), 2007, pp. 239-334. DOI : https://doi.org/10.1093/hwj/dbm040

[6] Plus amples informations sur le format IIIF (International Image Interoperability Framework, https://iiif.io/) suivront dans le rapport.

de l'iconographie ? Certaines images sont-elles représentatives de certains endroits ? Comment mesurer l'impact visuel d'une image ? La vaste quantité d'imprimés, revues artistiques, ou autres journaux de propagandes et leur variété témoigne d'un phénomène de mouvement à l'échelle globale. Les contagions visuelles, liant une iconographie à une autre, permettent d'établir des *visual block busters* amenant une réflexion internationale.

Ce rapport répond donc aux besoins quantitatifs et analytiques et constitue un socle pour l'élaboration de la carte interactive pour le site web. Dans un premier temps, seront exposées et décrites les données – ou le *data set*. Après quoi, nous proposerons un petit détour méthodologique mettant en exergue la marche à suivre et des indications que nous jugeons pertinentes pour la bonne utilisation de la plateforme. Dans un troisième temps, nous nous pencherons sur les résultats de l'analyse statistique qui sera postérieurement commentée dans la partie discussion. En annexe, seront proposés les notebooks Python nécessaires à l'expérience ainsi qu'une procédure d'utilisation de la plateforme.

Première partie – Méthodologie

La méthodologie adoptée dans le cadre de cette étude peut être divisée en deux parties : celle relative à la création de *clusters* (I), et celle relative à leur analyse (II).

I) L'étape de création de *clusters*

Dans le cadre du projet *Visual Contagions*, une plateforme, *Visual Contagions Explore (VCE)*, permet de regrouper entre elles, grâce à un algorithme de comparaison, des images qui partagent un certain niveau de similarité. Dans le cadre de cette étude, ce sont des images de revues qui ont été analysées par *VCE*.

L'objectif du projet est *in fine* de pouvoir détecter des circulations d'images à travers le temps, et l'espace ; et cela à un niveau d'échanges mondial. À cette fin, la nécessité de détenir un vaste ensemble d'informations relatives à chacune des images, nécessite l'utilisation de « manifestes ».

Un manifeste, c'est quoi ?

Un manifeste peut être comparé à une revue, puisqu'il permet de situer une image dans un *contexte*, une collection spécifique. Collection spécifique qui peut donc être une revue. L'avantage du manifeste, c'est que peut y être retrouvé un ensemble conséquent d'informations relatives à chacune de ses images. Au niveau informatique, le manifeste répond aux règles du format IIIF.

Format IIIF ?

L'*International Image Interoperability Framework* (IIIF) désigne ici un ensemble de spécifications techniques dont l'objectif est de définir un cadre d'interopérabilité pour la diffusion et l'échange d'images haute résolution sur le Web.[7] Le format IIIF est donc le format qui répond à ces spécifications.

[7] https://fr.wikipedia.org/wiki/International_Image_Interoperability_Framework

Il s'agit donc de *fournir VCE en manifestes*, en copiant-collant sur cette plateforme des URLs renvoyant à ces mêmes manifestes. Une fois ces manifestes fournis à *VCE*, cette plateforme y recherche des *clusters* (cf. définition ci-dessous).

Qu'entend-on par cluster *?*

Dans cet ensemble de données, nous définissons un « *cluster* » comme un groupe d'images qui ont été séparées du groupe général comme étant identiques ou presque identiques (une jauge de similarité a, à cet effet, été fixée à 0,92 ; sachant qu'une jauge fixée à 1 ne détecterait que des images strictement identiques entre elles). Nous verrons que, malgré cela, un certain nombre de *clusters* ont été formés avec des images seulement similaires, voire foncièrement différentes les unes des autres. D'où la présence, dans ce rapport, d'un travail de classification des *clusters*.

On désigne *via* substantivation, l'action de création de *clusters* par le terme de clusterisation. Cette *clusterisation* est réalisée par un algorithme d'apprentissage automatique.

II) L'analyse des *clusters*

Dans cette étude, les images réparties au sein de *clusters* sont dénommées : « image *clusterisées* ».

Les *clusters* détectés par la plateforme peuvent être étudiés directement sur le site Web de la plateforme[8] ou en les exportant sous la forme d'un fichier JSON créé par *VCE*.

Fichier JSON ?

JavaScript Object Notation (JSON) est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Il permet de représenter de l'information structurée.[9]

Dans le cadre de cette étude, lorsque cela était possible, les *clusters* ont systématiquement été exportés (téléchargés) ; et ce, dans l'optique de les traiter *via* de la programmation informatique en langage *Python*.

C'est grâce à des *Jupyter Notebooks* (dont le format est basé sur JSON), qui permettent, en alignant des lignes de code, de créer des programmes informatiques, qu'ont pu être réalisées une classification et une analyse de certains *clusters*.

[8] https://visualcontagions.unige.ch/explore/
[9] https://fr.wikipedia.org/wiki/JavaScript_Object_Notation

- Un premier *Jupyter Notebook* (cf. Annexe 2) permet la création d'un fichier CSV avec, pour chaque image clusterisée, les informations suivantes :

  - L'URL du manifeste d'où provient l'image *clusterisée* ;
  - L'URL de l'image *clusterisée* ;
  - L'URL de la page de la revue d'où provient l'image *clusterisée* ;
  - Le numéro du *cluster* dans lequel a été incluse l'image *clusterisée* ;
  - *Seulement en ce qui concerne les images* clusterisées *ayant pour origine Gallica : Le titre de la revue dans laquelle a été publiée l'image* clusterisée *;*
  - *Seulement en ce qui concerne les images* clusterisées *ayant pour origine Gallica : La date de publication du numéro de la revue dans lequel a été publiée l'image clusterisée.*

- Un second *Jupyter Notebook* (cf. Annexe 2) permet successivement :

  - La création d'un fichier CSV *fruit* d'une *fusion* entre le *classeur* d'origine (converti en CSV) d'où proviennent les URLs de manifestes fournies à *VCE*, d'une part, et le fichier CSV d'images *clusterisées* (créé grâce au premier *Jupyter Notebook*), d'autre part ; et ce, afin de réunir dans un même fichier les informations relatives aux *clusters* en eux-mêmes (numéros des *clusters*) et les informations d'ordre spatial relatives aux images *clusterisées* ;
  - La création d'un fichier CSV classant, dans un ordre de décroissant, les *clusters* en fonction de leur niveau de diffusion spatiale : c'est-à-dire, ici, leur nombre de villes dans lesquelles se retrouvent leurs images *clusterisées*.

- Un troisième *Jupyter Notebook* (cf. Annexe 2) permet une analyse des données notamment en :

  - Affichant les *clusters* avec le plus d'images (*clusterisées*) ;
  - Calculant la médiane des nombres d'images (*clusterisées*) par *cluster* ;
  - Calculant les moyenne et médiane des dates des publications des images *clusterisées* ;
  - Calculant les moyenne et médiane de n'importe quel type d'information souhaité relativement aux *clusters* et leurs images.

Ces différents fichiers CSV obtenus, une analyse des *clusters* est envisageable :

- Soit en commençant par analyser les *clusters* les plus *volumineux* (contenant le plus d'images) – et ce, en se fondant sur une corrélation : *cluster* volumineux *égal* plus de chance de découvrir une circulation d'une image ;
- Soit en commençant par analyser les *clusters* les plus *spatialement diffus* – et ce, en se fondant sur une corrélation : *cluster* spatialement très diffus *égal* plus de chance de découvrir une circulation d'une image.

Parallèlement à cette analyse ayant pour optique directe la recherche de circulations d'images, une classification des *clusters* (*clusters* exploitables / inexploitables) est opérée.

Deuxième partie – Description des données

ASPECTS LÉGAUX DU PROJET

Le projet Visual Contagions a nécessité l'analyse d'images afin d'établir des circulations d'images provenant de revues illustrées, imprimés et journaux de propagande à des fins de recherches légitimes. Dans ce cadre se pose la question du droit d'auteur et de la propriété intellectuelle.

Le projet a analysé les images par le biais de techniques dites du *data mining* et ainsi bénéficié, à l'égard des œuvres analysées encore sous droit d'auteur et sans licence permissive telles que les *Creative commons*, de l'exception de droit d'auteur "*d'utilisation d'oeuvres à des fins de recherche scientifique*" prévue à l'art. 24d LDA (appelée aussi exception de text and data mining "TDM").

La publication des résultats de l'analyse bénéficie par ailleurs, à l'égard des œuvres qui sont encore sous droit d'auteur et reconnaissables dans la publication, de l'exception de citation prévue à l'art. 25 LDA et toute autre éventuelle exception qui entrerait en ligne de compte, comme par exemple l'exception de l'inventaire prévue à l'art. 24e LDA.

La publication elle-même, co-réalisée entre le groupe d'étudiant-es n° 6 et l'Université de Genève, utilise la licence Creative commons BY et les résutlats sous-jacents la Creative commons 0 (p.ex. pour les données et métadonnées collectées et générées).

Ces informations sont fournies à titre informatif uniquement et ne doivent pas être interprétées comme des conseils juridiques.

Un ensemble de quatorze corpus et un corpus unifié

Cette étude repose d'une part sur un ensemble de quatorze différents corpus contenant (sauf exceptions) chacun des manifestes différents.

D'autre part, cette étude s'est appuyée sur un corpus unique, résultat de la fusion des quatorze corpus. Il sera fait référence à ce corpus-là sous l'appellation de *corpus unifié*.

Aussi bien pour l'ensemble de quatorze corpus que pour le corpus unifié, la période étudiée concerne les années comprises **entre 1920 et 1939.** Autrement dit, toutes les revues dont ont été étudiées les images ont été publiées, pour les plus anciennes en 1920, et pour les plus « récentes » en 1939.

Une description des données de l'ensemble de quatorze corpus (I) précédera celle du corpus unifié (II).

I) Description des données de l'ensemble de quatorze corpus

Les bases de données d'origine des manifestes de l'ensemble de quatorze corpus

L'ensemble de quatorze corpus est constitué d'environ deux mille magazines sous un format IIIF issus de trois institutions. La première est Gallica[10], la bibliothèque numérique de la Bibliothèque nationale de France. Les autres manifestes proviennent des archives numérisées de l'Université de Heidelberg[11], ainsi que de la bibliothèque numérique de la Bibliothèque nationale de Pologne : Polona[12].

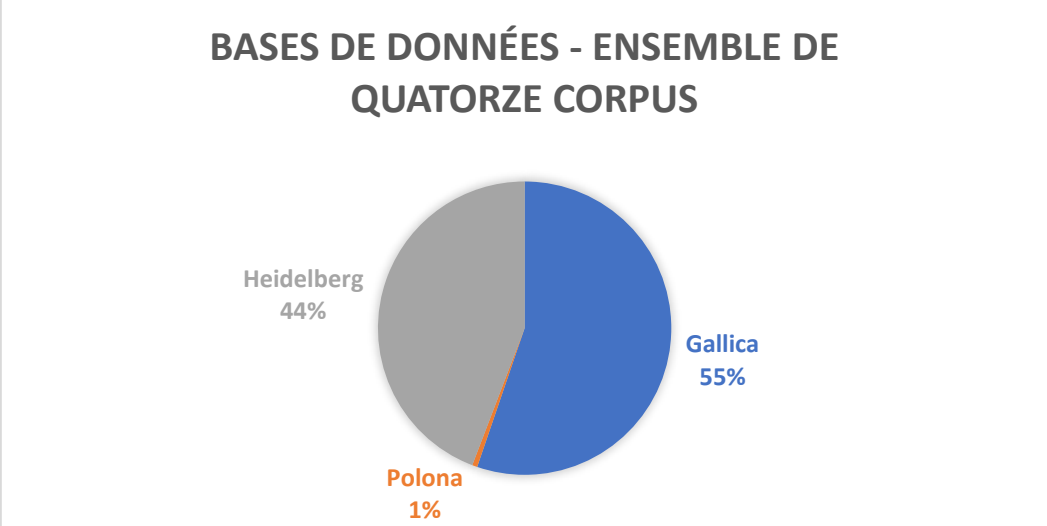
Ci-dessous, un tableau représentant la répartition des manifestes, pour chacun des quatorze corpus, selon leur base de données d'origine :

	Polona	Gallica	Heidelberg
corpus_0	6	94	0
corpus_1	0	200	0
corpus_2	0	192	8
corpus_3	0	0	200
corpus_4	0	0	126
corpus_5	0	58	72
corpus_6	0	58	72
corpus_7	0	5	4
corpus_8	0	5	4
corpus_9	0	7	243
corpus_10	0	6	96
corpus_11	0	250	0
corpus_12	6	163	0
corpus_13	0	136	114
Totaux	12	1 174	939

Répartition, pour chacun des quatorze corpus, des manifestes en fonction de leur base de données d'origine

La majorité des manifestes proviennent de Gallica, à hauteur de 55 % alors que Heidelberg et Polona se partagent les 45 % restants. Ainsi, au sein des 2 125 manifestes, seuls 12 proviennent de Polona, 1 174 de Gallica et 939 de Heidelberg.

Compte tenu des pays de rattachement des bases de données – respectivement la France, l'Allemagne et la Pologne, il est de prime abord constatable que le corpus constitué sera majoritairement nord-européen ainsi les circulations se concentreront dans cette spatialité. Il ne faut cependant pas adopter une approche limitative. Bien que les bibliothèques nationales comptent majoritairement des manifestes de leur pays, il est également envisageable que des excursions spatiales soient représentées sur les plateformes. De fait, Gallica offre, outre une base française – métropolitaine et outre-mer – colossale, une série de revues africaines, américaines, asiatiques ainsi qu'un corpus européen.



[10] https://gallica.bnf.fr/ ci-après : Gallica

[11] https://www.ub.uni-heidelberg.de/ ci-après : Heidelberg

[12] https://polona.pl/ ci-après : Polona

Répartition des manifestes en fonction des bases de données

Certaines URLs de manifestes de l'ensemble de quatorze corpus n'ont pas été « retenues » par la plateforme

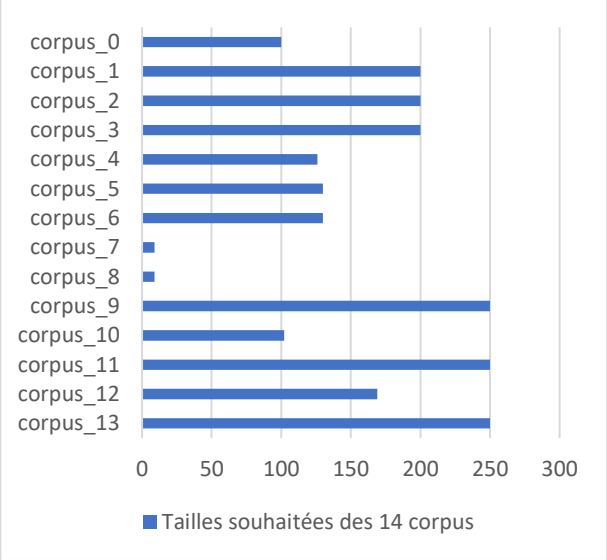
Parmi les URLs de manifestes utilisées pour la création de cet ensemble de quatorze corpus certaines n'ont pas pu être analysées par la plateforme. Lorsqu'il s'est agi de fournir la plateforme *VCE* (*Visual Contagions Explore*) en URLs de manifestes (ci-après : URLs), c'est-à-dire copier-coller des URLs de manifestes, d'un *classeur* (fichier XLSX) vers cette plateforme ; certaines URLs n'ont pas été « retenues » par *VCE*. Autrement dit, il est arrivé à de nombreuses reprises qu'après avoir copié-collé par exemple : 100 URLs sur *VCE*, seulement 98, par exemple, ne soient effectivement dans le corpus *VCE* créé à cette occasion.

Ci-dessous, des représentations du niveau de rétention/acceptation des URLs par *VCE*, pour l'ensemble de quatorze corpus :

URLs copiées-collées sur la plateforme *VCE* :

corpus_0	100
corpus_1	200
corpus_2	200
corpus_3	200
corpus_4	126
corpus_5	130
corpus_6	130
corpus_7	9
corpus_8	9
corpus_9	250
corpus_10	102
corpus_11	250
corpus_12	169
corpus_13	250
Total	2 125

URLs copiées-collées, dans chacun des quatorze corpus, sur VCE



Représentation de la « taille souhaitée » des quatorze corpus

URLs retenues par la plateforme *VCE* :

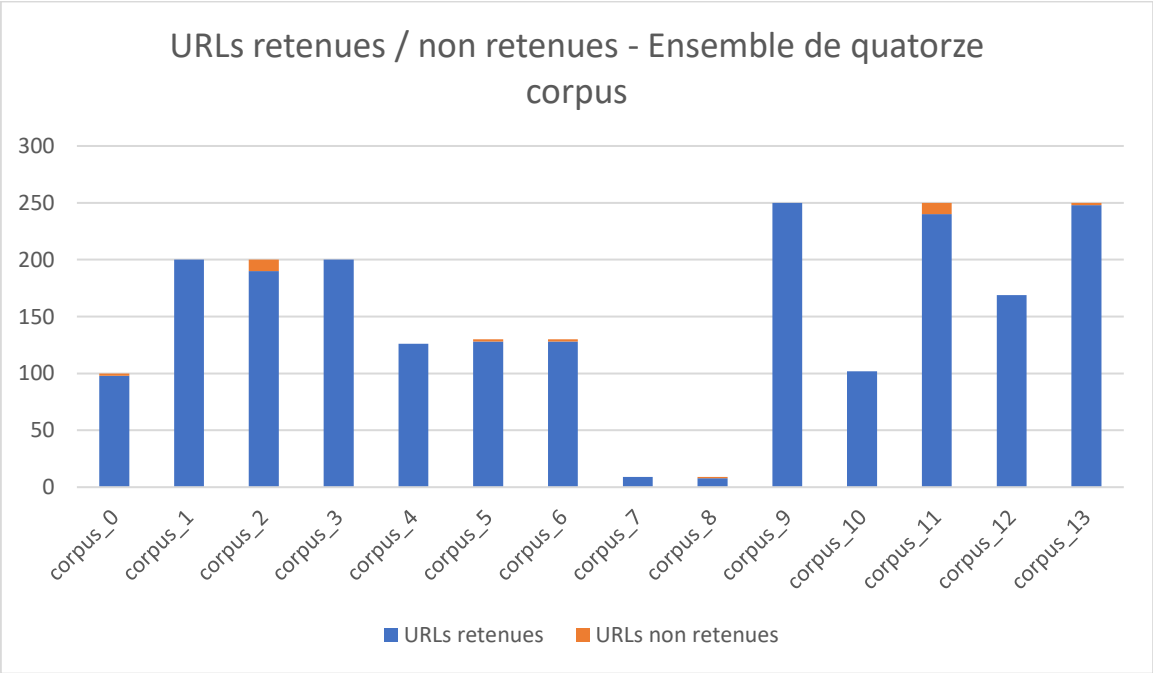
corpus_0	98
corpus_1	200
corpus_2	190
corpus_3	200
corpus_4	126
corpus_5	128
corpus_6	128
corpus_7	9
corpus_8	8
corpus_9	250
corpus_10	102
corpus_11	240
corpus_12	169
corpus_13	248
Total	2 096

URLs retenues par VCE *pour chacun des quatorze corpus*

Pourcentages d'URLs retenues :

corpus_0	98 %
corpus_1	100 %
corpus_2	95 %
corpus_3	100 %
corpus_4	100 %
corpus_5	98 %
corpus_6	98 %
corpus_7	100 %
corpus_8	89 %
corpus_9	100 %
corpus_10	100 %
corpus_11	96 %
corpus_12	100 %
corpus_13	99 %
% sur le total	99 %

Pourcentages, pour chacun des quatorze corpus, d'URLs retenues par VCE



Représentation, pour chacun des quatorze corpus, du taux de rétention/acceptation des URLs par VCE

Sur l'ensemble des quatorze corpus, **99 % des URLs ont été retenues**.

Certains manifestes de l'ensemble de quatorze corpus n'ont pas d'images

Parmi les manifestes retenus, certains ne comprenaient pas d'images exploitables : soit la plateforme n'a pas segmenté les images présentes ou la revue ne contenait pas d'images.

Ci-dessous, des représentations statistiques relatives à la présence ou non d'images au sein des manifestes de l'ensemble de quatorze corpus :

Manifestes avec images détectés par la plateforme *VCE* :

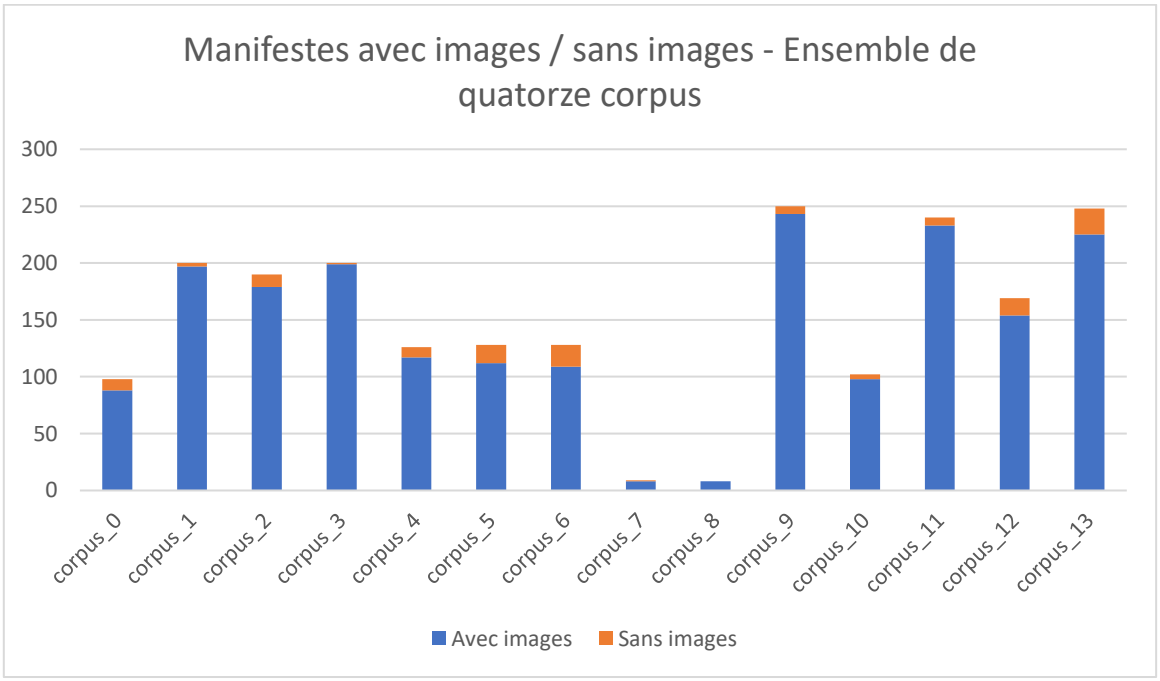
corpus_0	88
corpus_1	197
corpus_2	179
corpus_3	199
corpus_4	117
corpus_5	112
corpus_6	109
corpus_7	8
corpus_8	8
corpus_9	243
corpus_10	98
corpus_11	233
corpus_12	154
corpus_13	225
Total	1 970

Pour chacun des quatorze corpus et au total : manifestes avec images

Pourcentages de manifestes avec images :

corpus_0	90 %
corpus_1	99 %
corpus_2	94 %
corpus_3	100 %
corpus_4	93 %
corpus_5	88 %
corpus_6	85 %
corpus_7	89 %
corpus_8	100 %
corpus_9	97 %
corpus_10	96 %
corpus_11	97 %
corpus_12	91 %
corpus_13	91 %
% sur le total	94 %

Pour chacun des quatorze corpus et pour le total : pourcentages de manifestes avec images



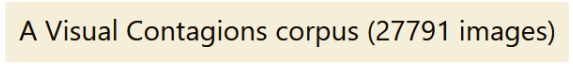
Représentation, pour chacun des quatorze corpus, du taux de manifestes – détectés par VCE *– contenant des images*

Sur l'ensemble des quatorze corpus, **94 % des manifestes comprenait des images**.

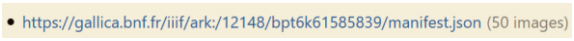
Les nombres d'images dans l'ensemble de quatorze corpus

Nombre d'images pour chacun des quatorze corpus :

corpus_0	26 115
corpus_1	4 104
corpus_2	27 791
corpus_3	90 230
corpus_4	42 723
corpus_5	42 028
corpus_6	42 028*
corpus_7	3 989
corpus_8	3 989*
corpus_9	93 029
corpus_10	49 072
corpus_11	12 085
corpus_12	9 054
corpus_13	74 679
Total	520 916 - 42 028** - 3 989** = **474 899**



Ce qu'affiche VCE *quant au nombre d'images d'un corpus*
(exemple : corpus_2)



Ce qu'affiche VCE *à la suite d'une URL quant au nombre d'images du manifeste auquel elle renvoie (exemple tiré du corpus_2)*

Ce qu'affiche VCE *à la suite d'une URL si le manifeste auquel elle renvoie ne contient pas d'images*

*doublons
**termes soustraits,
car relatifs à des images doublonnes

II) Description des données du corpus unifié

Le corpus unifié est issu d'une fusion de l'ensemble des quatorze corpus.

Nombre d'URLs copiées-collées sur le corpus unifié

1 970 URLs au sein de l'ensemble de quatorze corpus comprenaient des images. Grâce à un « nettoyage » préalable de ces quatorze corpus (c'est-à-dire y retirer les manifestes sans images) (cf. Annexe 1), il ne s'agissait dès lors plus que de copier-coller ces 1 970 URLs au sein d'un corpus unique : le corpus unifié.

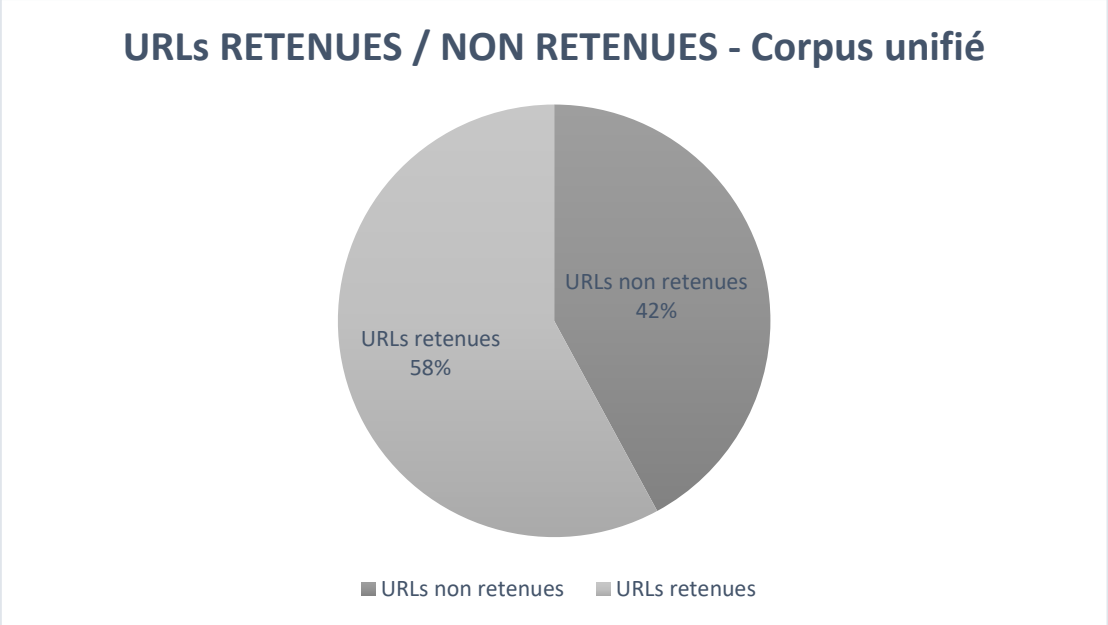
Néanmoins, s'est présenté un problème relatif à la présence de corpus doublons. En effet, le corpus_6 est un doublon du corpus_5 ; et le corpus_8 est un doublon du corpus_7. Lors de la fourniture d'URLs à la plateforme, une erreur (humaine) a été commise s'agissant du transfert des données du fichier *classeur* vers *VCE*. Mais, *VCE* reconnaît les URLs identiques et empêche qu'elles soient copiées-collées au sein d'un même corpus. C'est pourquoi, quand il s'est agi de copier-coller toutes les URLs relatives à des manifestes avec images de l'ensemble de quatorze corpus vers un seul corpus (le corpus unifié), ont été *de facto* copiées-collées non pas 1 970 URLs, mais : 1 970 - 109* - 8** = **1 853 URLs**.

*corpus_6
**corpus_8

1 853 URLs ont donc été copiées-collées sur le corpus unifié.

Nombre d'URLs retenues par VCE *au sein du corpus unifié*

Pour le corpus unifié, le nombre d'URLs retenues par *VCE* ne s'élève qu'à **1 073** (c'est-à-dire seulement 58 % environ des URLs copiées-collées !)



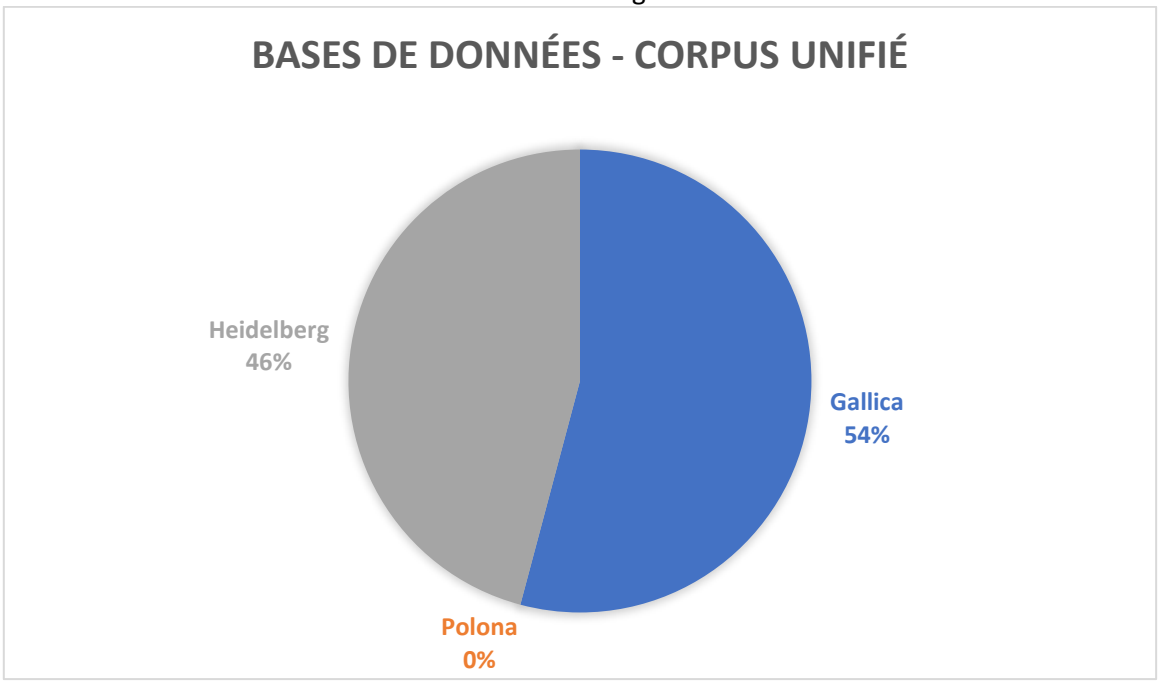
Répartition, pour le corpus unifié, *des URLs selon qu'elles aient été retenues ou non par* VCE

Bases de données d'origine des manifestes du corpus unique

Sur ce corpus unifié *d'URLs retenues par* VCE*, qu'en est-il des origines des manifestes ?*

Répartition des manifestes retenus par *VCE* selon leur base de données d'origine :

581 : Gallica 492 : Heidelberg 0 : Polona



Répartition des manifestes du corpus unifié *en fonction des bases de données*

Nombre d'images au sein du corpus unifié

Le nombre d'images au sein du corpus unifié s'élève à : **283 936** images ; c'est-à-dire presque moitié moins que les images de l'ensemble de quatorze corpus (cf. *supra*).

Troisième partie – Résultats

Seront présentés, dans un premier temps, les résultats de la *clusterisation* relative à l'ensemble de quatorze corpus (I), avant que ne soient présentés, dans un second temps, ceux relatifs à la *clusterisation* du corpus unifié (II).

I) Résultats pour l'ensemble de quatorze corpus

Pourquoi différents corpus ?

Dans un premier temps, c'est la démarche suivante qui avait été adoptée : ne créer qu'un seul corpus sur *VCE*, et lui ajouter des URLs de manifestes au fur et à mesure des repérages de *clusters* par la plateforme ; et ce, par « vagues » de 300 URLs maximum. (300 URLs maximum, car il s'agit de la capacité maximale de traitement de *VCE*.) Ainsi, issus du même corpus, auraient été récupérés plusieurs fichiers JSON (intitulés manifest.json par *VCE*) à la suite de chaque « vague ». L'avantage de cette méthode étant que le dernier fichier JSON qui aurait été récupéré aurait contenu tous les *clusters* relatifs à toutes les URLs du corpus unique.

Illustration :

1) Ajout dans le corpus unique de 300 URLs de manifestes ; le corpus unique contient maintenant 300 manifestes ; après quelques heures, *VCE* nous donne un fichier JSON contenant les *clusters* relatifs à ces 300 manifestes.
2) Ajout dans le corpus unique de 300 URLs de manifestes ; le corpus unique contient maintenant 600 manifestes ; après quelques heures, *VCE* nous donne un fichier JSON contenant les *clusters* relatifs à ces 600 manifestes. La capacité maximale de traitement de *VCE* n'est pas dépassée, car les 300 premiers manifestes ont déjà été traités.
3) Etc.
4) Ajout dans le corpus unique de 300 URLs de manifestes ; le corpus unique contient maintenant x manifestes (ex : 4 500) ; après quelques heures, *VCE* nous donne un fichier JSON contenant les *clusters* relatifs à ces x manifestes. La capacité maximale de traitement de *VCE* n'est pas dépassée, car les (x - 300) premiers manifestes ont déjà été traités.

Cependant, dès la seconde « vague » nous avons constaté des problèmes sur la plateforme (échecs répétés à obtenir des *clusters*, beugs) nous ayant amenés, inévitablement, à annuler des tâches données à la plateforme – ce qui était déconseillé.

C'est la raison pour laquelle, il a été décidé de créer un corpus pour chaque « vague », et, à la fin, de fusionner tous ces corpus en un seul : le corpus unifié.

Clusters *découverts par la plateforme* VCE *dans l'ensemble de quatorze corpus*

Ci-dessous, deux tableaux présentant, en ce qui concerne l'ensemble de quatorze corpus, les nombres de *clusters* trouvés par *VCE*, et les moyennes des nombres de *clusters* par manifeste.

Clusters trouvés par la plateforme *VCE* :

corpus_0	1 336
corpus_1	115
corpus_2	1 370
corpus_3	3 214
corpus_4	1 181
corpus_7	145
corpus_8	145
corpus_9	3 061
corpus_10	1 896
corpus_13	2 857
Total	**15 320**

Moyennes de *clusters* par manifeste :

corpus_0	26,72
corpus_1	0,58
corpus_2	7,65
corpus_3	16,15
corpus_4	10,09
corpus_7	18,13
corpus_8	18,13
corpus_9	12,6
corpus_10	19,35
corpus_13	12,7
Moy. sur le total	14,21

Remarque : Nous ne disposons pas de résultats pour les corpus 5, 6, 11 et 12 en raison d'une beug de la plateforme.

La moyenne sur l'ensemble des *clusters* (non les valeurs « corpus ») s'élève à **11,25** *clusters* par manifeste.

Les nombres d'images clusterisées *au sein de l'ensemble de quatorze corpus*

Remarque : Parmi les images identifiées dans chaque manifeste une partie seulement sera prise en compte dans la *clusterisation*. De sorte qu'il faut, afin d'éviter toute méprise, distinguer « images des manifestes » et « images *clusterisées* ».

Ci-dessous, un tableau présentant les nombres d'images *clusterisées* à partir de l'ensemble de quatorze corpus.

Nombres d'images *clusterisées* :

corpus_0	5 927
corpus_1	553
corpus_2	7 501
corpus_3	45 226
corpus_4	15 836
corpus_7	498
corpus_8	498
corpus_9	43 829
corpus_10	19 097
corpus_13	28 006
Total	**166 971**

Moyennes et médianes des nombres d'images par cluster *de l'ensemble des quatorze corpus*

Ci-dessous, un tableau représentant les moyennes et les médianes d'images par *cluster* de l'ensemble de quatorze corpus. Les médianes peuvent notamment être obtenues grâce à des lignes de code *Python* (cf. Annexe 2) :

	Moyennes d'images par *cluster*	Médianes d'images par *cluster*
corpus_0	4,44	2
corpus_1	4,81	2
corpus_2	5,48	2
corpus_3	14,07	2
corpus_4	13,41	2
corpus_7	3,43	2
corpus_8	3,43	2
corpus_9	14,32	2
corpus_10	10,07	2
corpus_13	9,8	2
Total	8,33	2

Il est à noter que lorsque l'on considère non pas les valeurs « corpus » mais plutôt la valeur « somme de tous les *clusters* », la moyenne d'images par cluster ne s'élève non pas à 8,33 mais à **10,9 images par *cluster*.**

II) Résultats pour le corpus unifié

Clusters *découverts par la plateforme* VCE *dans le corpus unifié*

VCE a détecté **11 609** ***clusters*** au sein du corpus unifié.

Nombre d'images clusterisées *au sein du corpus unifié*

120 199 images ont été ***clusterisées***.

Bases de données d'origine des images clusterisées *au sein du corpus unifié*

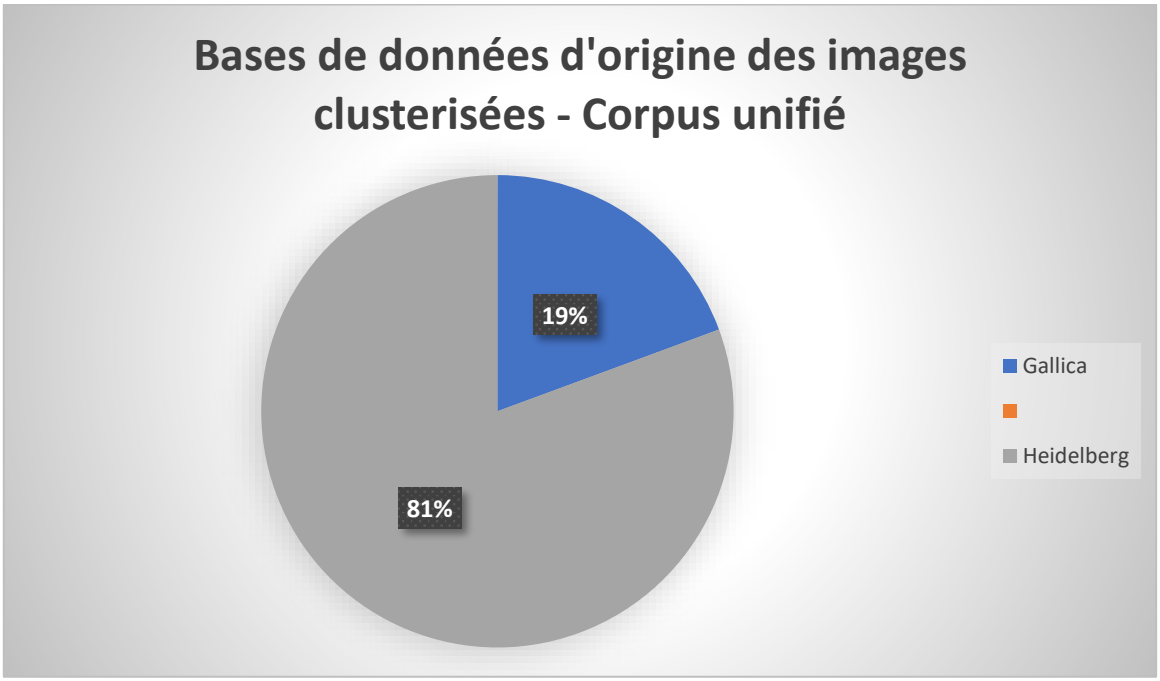
Il est intéressant d'observer qu'alors que c'était Gallica qui comptabilisait le plus grand nombre d'URLs retenues pour ce corpus unifié (54 % – cf. *supra*), c'est Heidelberg qui détient, pour ce même corpus unifié, le plus grand nombre d'images *clusterisées* :

Nombres d'images *clusterisées* pour chaque base de données :

97 553 de Heidelberg

23 446 de Gallica

(Rappel : 0 de Polona, car 0 URL retenue pour ce corpus unifié.)



Répartition des images clusterisées *du corpus unifié selon leur base de données d'origine*

Moyenne et médiane des nombres d'images par cluster *du corpus unifié*

Moyenne d'images par *cluster* :

10,42

Médiane des images par *cluster* (obtenue grâce à des lignes de code *Python* (cf. Annexe 2)) :

2

Quatrième partie – Analyse des résultats

Les résultats statistiques de cette étude nous ont amenés à établir une classification partielle des *clusters* de l'ensemble de quatorze corpus (II), ainsi qu'une analyse (couplée à une classification partielle également) des *clusters* du corpus unifié (III). Avant cela, il s'était agi d'effectuer une présentation des différentes catégories de *clusters* envisageables (I).

I) Les catégories de *clusters* inexploitables

Classifier les *clusters* – Avant de présenter une classification partielle des *clusters* obtenus grâce à *VCE*, il s'agit de présenter les raisons nécessitant une telle classification. En effet, il a été décidé d'axer l'étude des *clusters*, au-delà de la seule recherche de circulations d'images (de contagions visuelles), sur le repérage de *clusters* inexploitables aux fins d'une telle recherche.

En somme, l'identification d'une variété de *clusters* a appelé une tentative de classification de ces derniers. Cependant, cette classification n'a pu concerner qu'une partie infime des *clusters*.

Cela étant dit, il faut préciser qu'une telle classification, même partielle voire très partielle, pourra être utile à toute personne souhaitant poursuivre la recherche de circulations d'images sur un ou des corpus similaires.

En effet, toute étude statistique ultérieure bénéficierait d'une telle classification préalable approfondie des *clusters*. Et ce, notamment s'agissant des stratégies à adopter face aux données afin de tenter de quantifier d'éventuelles circulations.

C'est un cas de *cluster* issu de l'ensemble de quatorze corpus qui permettra d'illustrer la première catégorie des *clusters* que nous qualifions de « stériles » (A).

Au sein de la seconde catégorie : les *clusters* d'images non identiques (notamment illustrée par un cas de *cluster* du corpus_2) une distinction peut être faite entre les *clusters* d'images similaires (A), et les *clusters* incohérents (B).

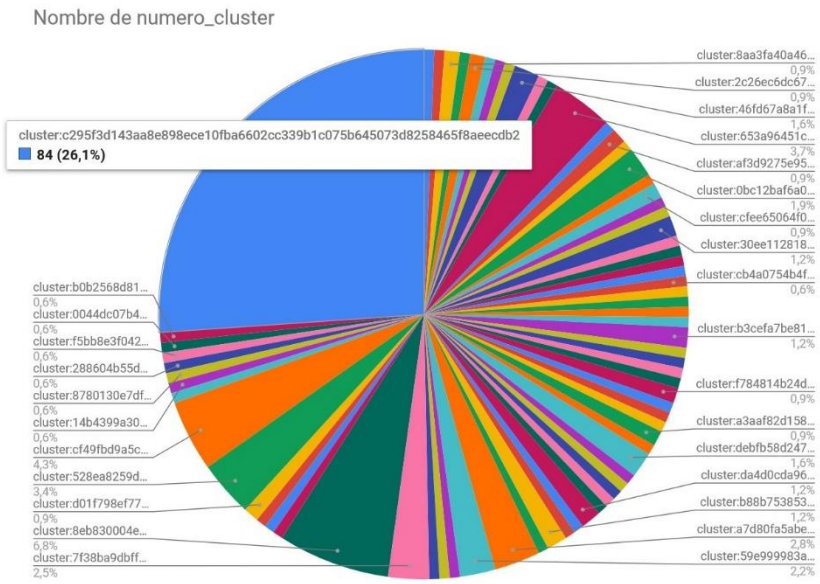
A) Première catégorie de *clusters* « inexploitables » : les *clusters* « stériles »

Ci-dessous, est présenté un cas illustratif du caractère « inexploitable » – dans le cadre précis de notre étude – d'un nombre considérable de *clusters*, car « stériles » dans une optique de recherche ou de reproduction de recherches de circulations d'images :

Il s'agit du cas d'un *cluster* issu du corpus_1, dans lequel *VCE* a notamment détecté 73 *clusters*, et à cette occasion, *clusterisé* 322 images.

Cas du *cluster* le plus volumineux du corpus_1

Quels étaient les nombres d'images dans les différents clusters *du corpus_1 ?*



Répartition des clusters *trouvés par* VCE *au sein du corpus_1, selon leur nombre d'images*

Le *cluster* le plus volumineux contenait 84 images, ce qui équivalait à environ 26 % des images *clusterisées* à partir du corpus_1 (*cluster* intitulé par *VCE* : cluster:c295f3d143aa8e898ece10fba6602cc339b1c075b645073d8258465f8aeecdb2).

Qu'a-t-il été découvert après observation du fichier CSV des images clusterisées *?*

A1 | manifeste_url

	A	B	C	D	E	F	G
1	l	image_url	page_url	ter	titre	date	type
2	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-janv-20	revues
3	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-janv-20	revues
4	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-févr-20	revues
5	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-févr-20	revues
6	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-mars-20	revues
7	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-mars-20	revues
8	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-avr-20	revues
9	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-avr-20	revues
10	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-mai-20	revues
11	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-mai-20	revues

Toutes les images de ce *cluster* provenaient de la même revue : *Le Bulletin de la vie artistique*, une revue qui semblait être, de toute évidence, bimensuelle.

76	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-déc-22	revues
77	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-janv-23	revues
78	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-janv-23	revues
79	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-janv-23	revues
80	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-févr-23	revues
81	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-févr-23	revues
82	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-mars-23	revues
83	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-mars-23	revues
84	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	01-avr-23	revues
85	https://galli	https://galli	https://galli	3d143aa8e8	de la vie	15-avr-23	revues
86							

Un classement des images dans un ordre chronologique permettait d'observer qu'il y avait (sauf quatre exceptions[13]) une date par image. Malgré le fait que les images provenaient de la même revue, une analyse d'une éventuelle circulation purement temporelle – sans considérations spatiales : la revue, la ville et le pays étant uniques – aurait pu *a priori* être envisagée. Cependant, il a été observé que ces 84 images provenaient toutes des couvertures de la revue *Le Bulletin de la vie artistique*. En l'espèce, toutes les couvertures possédaient le même logo : c'était la raison pour laquelle ce *cluster* s'était formé. Cela était clairement observable simplement par visualisation de quelques-unes des images de ce *cluster*.

[13] Deux images pour : 15 déc. 1920 ; 1er janv. 1921 ; 1er janv. 1922 ; 1er janv. 1923

Visualisation de deux de ces images :

La 1re, en date du 1er janvier 1920 : La 84e, en date du 15 avril 1923 :


Elles étaient identiques.

Elles provenaient chacune de deux couvertures de deux numéros différents de la revue *Le Bulletin de la vie artistique* :

La couverture du 1er janvier 1920 : La couverture du 15 avril 1923 :

En définitive, aucune circulation ne pouvait être observée ici. Surtout, ce cas-là indiquait que la situation d'images *clusterisées* en raison de leur présence sur la couverture d'une revue avait vocation à concerner potentiellement un nombre considérable de revues, voire à se répéter systématiquement. D'où l'utilité, dans une optique classificatrice, d'adjoindre l'épithète « stérile » à ce type de *cluster*. En effet, la recherche de circulations d'images ne saurait trouver un environnement fertile au sein de ces *clusters*, qui témoignent seulement d'itérations automatiques (ex : même logo sur des couvertures d'une revue), et non de quelque contagion visuelle.

De manière analogue à celle du cas des couvertures de revues, sont apparus ceux des quatrièmes de couvertures, des lettrines, des publicités, etc.

B) Seconde catégorie de *clusters* « inexploitables » : les *clusters* d'images non identiques

Le syntagme « non identique » recoupe ici les cas de *clusters* assemblant des images seulement « similaires », d'une part (première sous-catégorie) ; et les cas de *clusters* assemblant des images foncièrement différentes les unes des autres, d'autre part (seconde sous-catégorie) – ces derniers cas concernent des *clusters* qui seront réunis sous l'appellation de *clusters* « incohérents ». Il est à noter qu'un *cluster* « incohérent » est généralement seulement incohérent pris dans son ensemble ; de sorte qu'il assemble en partie des images similaires entre elles (voire identiques) mais aussi des images foncièrement différentes les unes des autres.

Exemple pratique : cas du cluster *le plus volumineux du corpus_2*

Ci-dessous, est présenté un cas illustratif du caractère inexploitable, dans le cadre de cette étude, d'un nombre considérable de *clusters*, car assemblant des images non identiques :

Il s'agit du cas du *cluster* le plus volumineux (1 069 images), issu du corpus_2, dans lequel *VCE* a détecté 1 370 *clusters*, et à cette occasion, *clusterisé* 7 501 images. Ce *cluster*, pris dans son ensemble, peut être qualifié d'incohérent.

Cas du *cluster* le plus volumineux du corpus_2


Les deux images du haut ne sont pas identiques (elles ne sont même pas *quasi* identiques) mais elles sont *similaires* au sens large du terme.
Celle du bas semble, de toute évidence, foncièrement différente.

Trois images extraites du cluster *:*
cluster:d6abcbffb4c9d193c1664a00cbed4e04cd39828497c19016bfcf4b77fe23b33f

Présentation des deux sous-catégories de *clusters* d'images non identiques :
Au sein de la seconde catégorie – *clusters* d'images non identiques – il s'agit de faire une différence entre :

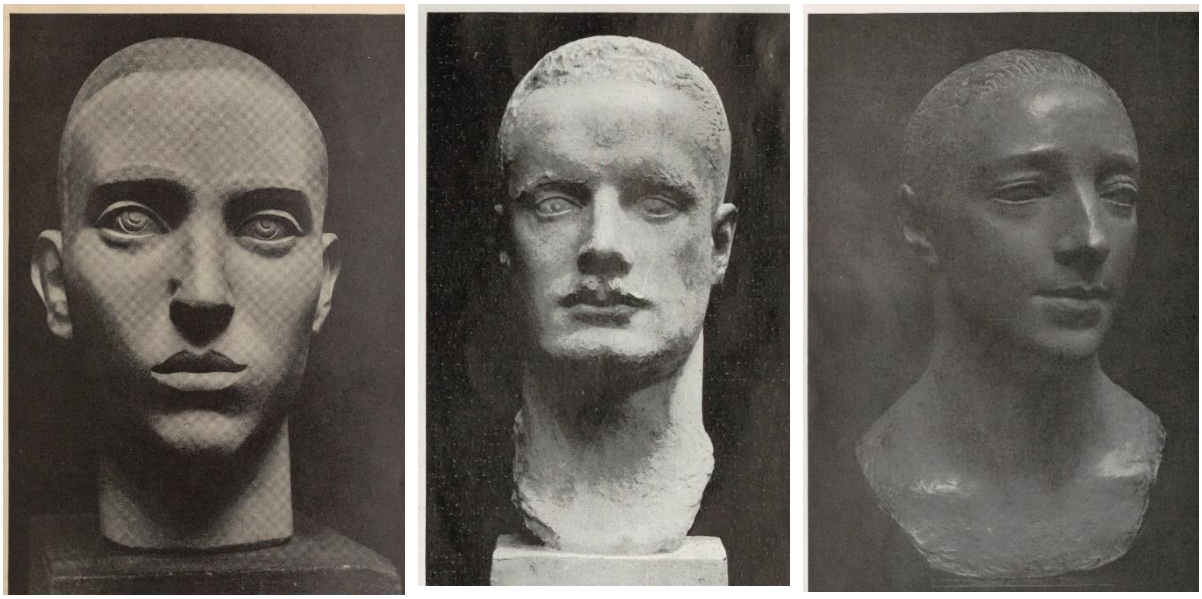
- *Clusters* d'images non identiques mais similaires dans une certaine mesure (ci-après : **première sous-catégorie**) ;

Et

- *Clusters* d'images incohérents (ci-après : **seconde sous-catégorie**).

1) Première sous-catégorie : les *clusters* d'images similaires

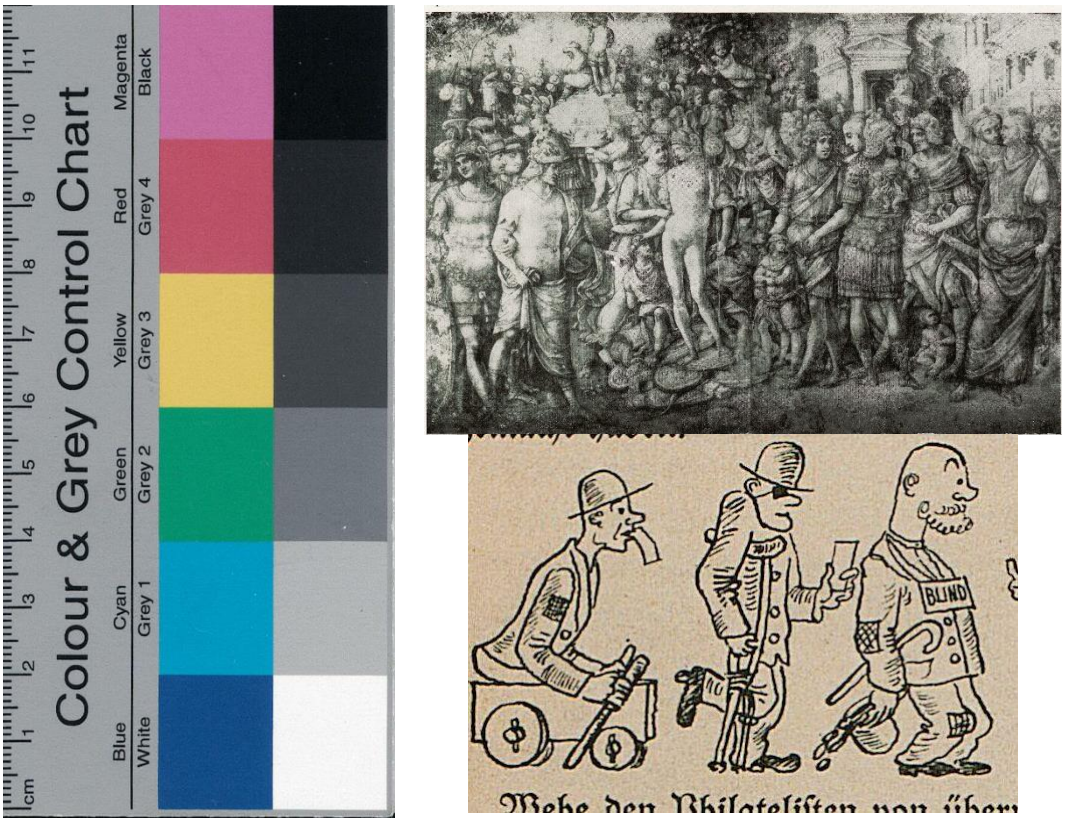
La première sous-catégorie de *clusters* peut être intéressante pour un autre type d'étude, mais pas dans le cadre de cette étude qui vise à observer, si l'on considère un élément seul, la circulation d'une image : cette image étant toujours la même (ou presque) le long de son chemin de diffusion. Cela étant dit, l'œil du profane ne peut être qu'ébloui par la capacité d'une intelligence artificielle à regrouper des images similaires (portraits, paysages, objets longilignes, etc.) ainsi que le ferait un être humain. Voici un exemple d'images issues de ce type de *clusters* :



Trois images extraites du cluster *:*
cluster:00005d008a899e179849e92791e25ae9de71d7d2eecae23f1b558df80b9f2bbf3feee2f2
Cluster *issu du corpus* unifié

2) Seconde sous-catégorie : les *clusters* « incohérents »

La seconde sous-catégorie de *clusters* illustre les aléas inévitables que présente le type de projets dont fait partie cette étude ; particulièrement dans le domaine de l'intelligence artificielle (type : réseau de neurones artificiels). En effet, l'incohérence de ces *clusters* semble évidente :



Trois images extraites du cluster *:*
cluster:01383900567c7733f03e58712b3f832f531935f4e190eeb2da0d646f38fab10bb872f2b6
Cluster *issu du corpus* unifié

Il est à noter que ces derniers *clusters* semblent assez rares, mais qu'ils peuvent contenir un nombre considérable d'images : d'aucuns diront qu'ils jouent un rôle de « fourre-tout ». (Les images au-dessus proviennent d'un *cluster* contenant plus de 100 000 images ; des images très diverses.)

Dans quel ordre a été opérée la classification partielle des clusters *dans l'ensemble de quatorze corpus ?*

Ensemble de quatorze corpus – classification des *clusters* les plus volumineux
En la matière, volumineux signifie : avec beaucoup d'images. C'est en raison de la corrélation : *cluster* volumineux *égal* plus de chance de découvrir une circulation d'images, qu'il a été décidé de classifier les *clusters* dans un ordre décroissant, en commençant par le plus volumineux. Cette corrélation est une corrélation postulée, préalablement à toute analyse des *clusters*, qui constituait, et constitue toujours dans une certaine mesure, l'un des fondements de travail de recherche de circulations d'images, dans le cadre de cette étude.

Dans quel ordre a été opérée la classification des clusters *du corpus unifié ?*

Corpus unifié – classification des *clusters* les plus spatialement diffus
S'agissant du corpus unifié, la classification des *clusters* – également partielle (*très partielle* : 16 sur 11 609), a été réalisée en commençant par les *clusters* les plus *spatialement diffus*. Il est question des *clusters* contenant des images retrouvées dans le plus de villes. En la matière, a été suivie une autre corrélation : *cluster* disposant d'images publiées dans beaucoup de villes *égal* plus de chance de découvrir une circulation d'images. Cette corrélation est également une corrélation postulée préalablement à toute étude des *clusters*, mais dont il est difficile de contester *a priori* le bien-fondé.

II) Classification partielle des clusters dans l'ensemble de quatorze corpus

Ci-dessous, une classification des *clusters* les plus volumineux des corpus_1, corpus_2 et corpus_7 issus de l'ensemble de quatorze corpus.

Remarque : Cette classification ne concerne que trois des quatorze corpus, pour des raisons tenant à des considérations organisationnelles. En effet, la possibilité d'obtenir un corpus unifié était soumise à la résolution d'un problème informatique. Ce n'est que tardivement, dans le cadre des limites de temps fixées pour cette étude, que ce dernier problème fut résolu. Dès lors, c'est la classification des clusters du corpus unifié qui était devenue prioritaire.

Classification des cinq *clusters* les plus volumineux du corpus_1 (73 *clusters*) :

Remarque : Il s'agit ici du premier jeu de clusters trouvés par VCE (73 clusters trouvés). Le second jeu de clusters trouvés par VCE au sein de ce corpus n'a pas fait l'objet d'une classification.

1er *cluster* : 84 images — ***cluster* « stérile »**
cluster:c295f3d143aa8e898ece10fba6602cc339b1c075b645073d8258465f8aeecdb2
COUVERTURES de la revue *Le Bulletin de la vie artistique*

2e *cluster* : 22 images — ***cluster* « stérile »**
cluster:8eb830004e9ab5af5ace831dd5860c603df9214cc1e868b4dd9aadcba2ac5622
QUATRIÈMES DE COUVERTURE de la revue *Transition / [dir. Eugene Jolas]*

3e *cluster* : 14 images — ***cluster* d'images similaires**
cluster:cf49fbd9a5c07bcf271f216501ca2b362d60287c8fed4a9837f70b0a778f1ab4
PAYSAGES

4e *cluster* : 12 images — ***cluster* « stérile »**
cluster:653a96451c3f68426e1fd8a5a677821a82625e745614ca24fca5c23be935bfa8
BANDEAUX BLANCS

5e *cluster* : 11 images — ***cluster* « stérile »**
cluster:528ea8259d4f9ede00d0e6cd4600020ceac374b2753ac0d024e616357f4d387b
QUATRIÈMES DE COUVERTURE de la revue *Transition / [dir. Eugene Jolas]*

Classification des cinq *clusters* les plus volumineux du corpus_2 (1 370 *clusters*) :

1er *cluster* : 1 069 images — ***cluster* « incohérent »**
cluster:d6abcbffb4c9d193c1664a00cbed4e04cd39828497c19016bfcf4b77fe23b33f

2e *cluster* : 375 images — ***cluster* d'images similaires ←→* *cluster* « incohérent »**
cluster:cc2ec929b7c76ef3499128d3f9b118b8cfe1745c23e1968ba64a98e97ddc261b
Paysages et, si présence d'individus, *plans larges* ; présence *quasi* systématique de végétation (arbres, buissons, etc.)

*non classifiable (*entre-deux*)

3e *cluster* : 360 images — ***cluster* d'images similaires ←→ *cluster* « incohérent »**
cluster:39f7f3668f27143c98ff2f6ac18478b4dc6e45608d854fb7345863d98a2977f6
Dessins avec des lignes *arrondies* marquées ; forte présence d'oiseaux en vol

4e *cluster* : 149 images — ***cluster* « stérile »**
cluster:38aee9e1e578bc036c32d641c4b14ea20c020a3a6a5c3b639e24edf946218b7d
COUVERTURES de la revue *Le Moniteur du dessin, de l'architecture & des beaux-arts : […]*

5e *cluster* : 136 images — ***cluster* « stérile »**
cluster:136d948061d1287aece12f42f1b0b593f8f5497b6c24b175c5e613d43cfd1eff
PUBLICITÉ

Classification des cinq *clusters* les plus volumineux du corpus_7 (145 *clusters*) :

1er *cluster* : 19 images — ***cluster* « stérile »**
cluster:bbe1996ae5b6d29fe3fbe1140afda409d8d4fa2919e442eb88315813a2e598af
PUBLICITÉ

2e *cluster* : 17 images — ***cluster* « stérile »**
cluster:f49f02d02fcb595b4c2f95ba45a9477ef2bf8cdae33e7d171d71dcf4f6c32275
LETTRINES

3e *cluster* : 13 images — ***cluster* « stérile »**
cluster:5b6cfbd8eb6db3c658ddc2c03d73a2ca37117bf100933aab202a6c06f3db382a
Photographie d'une œuvre du sculpteur Antoine Bourdelle en haut d'un bulletin de souscription afin d'obtenir l'édition de son œuvre complet ; dans la revue *L'Amour de l'art : […]*

4e *cluster* : 12 images — ***cluster* « stérile »**
cluster:83d36330e94f89143adbce154db223f4360aa77aa1351eb6303de7fbe593b601
PUBLICITÉ

5e *cluster* : 12 images — ***cluster* « stérile »**
cluster:71604e9d80120739b92c2943d181b2c73abb5be5f0eb2126fb088c20d480dbd4
PUBLICITÉ

III) Classification et analyse des *clusters* du corpus unifié

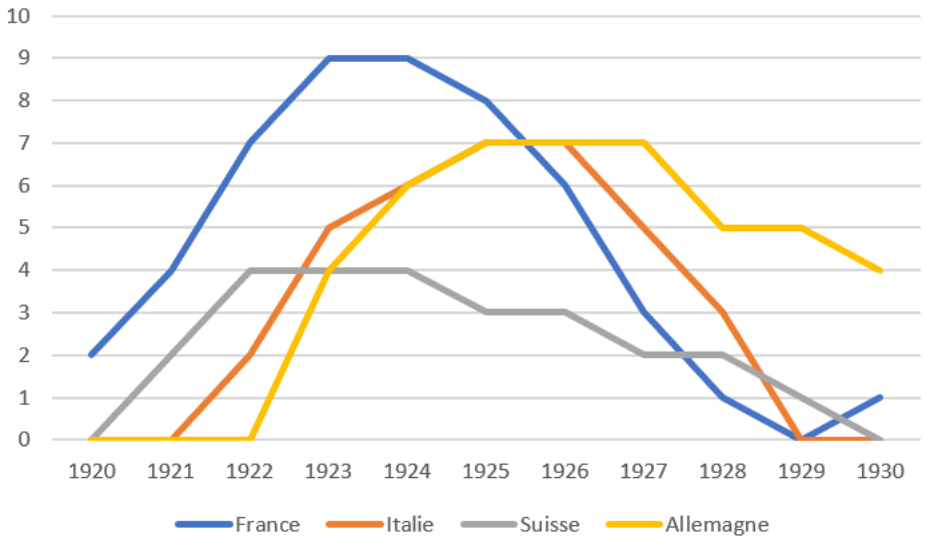
A) Mesurer la diffusion spatiale des images

Rechercher une ou des circulations d'images nécessite de prendre en considération, en corrélation avec le paramètre temporel, le paramètre de diffusion spatiale. En effet, une telle recherche permet – potentiellement – d'aboutir à la découverte d'une circulation « complète », en ce sens que cette dernière témoignerait de la présence à différents endroits (villes, pays, etc.) d'une image, dont la présence pourrait être *spatialement* et *temporellement variée*. D'où l'idée d'analyse d'une circulation « complète » (ou d'analyse « complète » d'une circulation) : faisant intervenir les variables *espace* et *temps*.

Analyser des circulations de manière « incomplète » reviendrait à soit ne considérer que le niveau de diffusion spatiale d'une image à un instant *t* ; soit ne considérer que le niveau de diffusion temporelle d'une image au sein d'un espace unique.

Au contraire, l'analyse « complète » d'une circulation rendrait, éventuellement, possible le fait de pouvoir *remonter le fil* d'une circulation d'une image. Exemple : Telle image se retrouve dans telle ville *a* à telle époque ; cette même image se retrouve dans telle autre ville *b* à une époque antérieure → le principe de causalité fait que, si circulation il y a, elle ne peut qu'aller dans un sens ville *b* vers ville *a*.

Ci-dessous, une illustration – avec un exemple fictif – de ce qui serait attendu avec des données adéquates :



Ces courbes fictives permettent d'envisager l'existence d'une circulation d'une image présente au sein d'un même cluster*, et retrouvée dans différents pays, sur différentes années (sur l'axe des ordonnées : le nombre d'images)*

Concernant les données de l'étude

Absence d'information d'ordre spatial dans les fichiers CSV – Cette étude a dû faire face à un écueil principal : lors de la création des fichiers CSV (ayant pour source aussi bien l'ensemble de quatorze corpus que le corpus unifié) avec les informations quant aux *clusters*, ne pouvait être incluse, dans ces fichiers CSV, aucune information d'ordre spatial – pour des raisons tenant à l'organisation des métadonnées des manifestes.

Résolution du problème pour le corpus unifié – Cet écueil a été surmonté par une fusion (obtenue grâce à des lignes de code *Python* (cf. Annexe 2)) du *classeur* (fichier CSV), d'où étaient prélevées les URLs à fournir à la plateforme *VCE*, avec le fichier CSV d'images *clusterisées* issu du corpus unifié.

Illustration de cette fusion des données :

1) Données issues du *classeur* d'origine avec les informations relatives notamment à la revue d'origine, le pays, la date et la ville ; mais bien sûr sans le *cluster* – à ce niveau, aucune *clusterisation* n'a été réalisée :

2) Données issues du fichier CSV des images clusterisées avec bien sûr le *cluster* (à ce niveau, la *clusterisation* a été réalisée), mais aucune information d'ordre spatial :

3) Données issues du fichier CSV *post*-fusion entre le *classeur* (fichier CSV également) et le fichier CSV des images *clusterisées*, avec notamment le *cluster*, la revue d'origine, la date, le pays, la ville :

Une fois les informations quant aux *clusters*, les informations quant aux revues d'origine des images, et les informations d'ordres temporel et spatial rassemblées dans un même fichier CSV, la recherche de circulations d'images – à travers *le temps* et *l'espace* – était dorénavant envisageable.

Au sein de ce fichier CSV intitulé : fusion_clusters_lieux_corpusMERGED_11609.csv, il ne restait plus qu'à classer les *clusters* selon leur niveau de diffusion spatiale.

Cette opération a été effectuée en s'axant sur l'information d'ordre spatial : « *City* » (c'est ainsi qu'elle était nommée dans le *classeur* d'origine et donc *in fine* dans le fichier CSV fusionné) ; c'est-à-dire l'information de la ville d'origine de l'image – entendre : la ville de publication de la revue d'où provient l'image.

Pourquoi prendre en considération l'élément : ville*, et non pas l'élément :* pays *?*

Tout simplement car, lorsqu'il s'agit de quantifier une diffusion spatiale, plus l'on réduit l'échelon d'analyse (en l'occurrence, passer de l'échelon *pays* à l'échelon *ville*), plus l'on augmente – *a priori* – les chances de découverte d'une telle circulation (et donc, autrement dit, plus l'on augmente les chances que la quantification de cette diffusion spatiale ne soit pas nulle).

Dans un premier temps, cela a permis d'aboutir (grâce à des lignes de code *Python* (cf. Annexe 2)) au classement suivant :

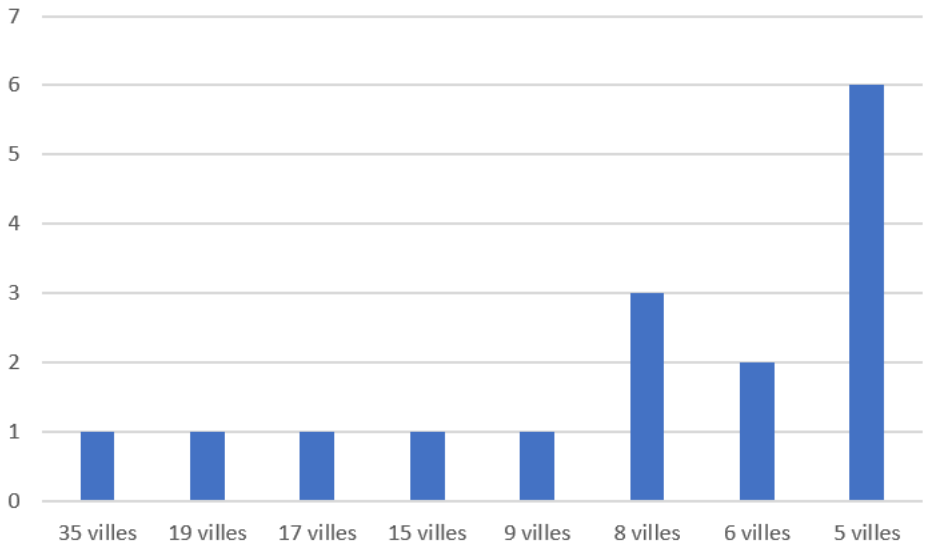
numero_cluster	City	image_url
cluster:01383900567c7733f03e58712b3f832f531935...	35	79929
cluster:000085003eb5ab4015e3c380232c0425651295...	19	133
cluster:0000c500515dfef4c36d6dc3788e03b3f1cae5...	17	197
cluster:00004f006b5c5e7922a38df995911d3d390b3c...	15	79
cluster:00003e005d8b310b66d7bb86c9e5c9c64b061e...	9	62
...	...	...

Remarque : En raison d'une « maladresse » lors de la fusion des données, la colonne « image_url » désigne ici le nombre d'images présentes au sein du cluster.

Dans un second temps, a pu être créé (grâce à des lignes de code *Python* (cf. Annexe 2)) un fichier mesurant le niveau de diffusion spatiale pour chaque *cluster* – c'est-à-dire le nombre de villes dans lesquelles se retrouvent les images d'un *cluster* donné. Cela a permis d'aboutir au classement suivant quant au corpus unifié :

> Relativement aux *clusters* ayant des images retrouvées dans un nombre de villes compris entre 35 et 5 :[14]
>
> Nombre de *clusters* ayant des images se retrouvant dans 35 villes : 1
>
> Nombre de *clusters* ayant des images se retrouvant dans 19 villes : 1
>
> Nombre de *clusters* ayant des images se retrouvant dans 17 villes : 1
>
> Nombre de *clusters* ayant des images se retrouvant dans 15 villes : 1
>
> Nombre de *clusters* ayant des images se retrouvant dans 9 villes : 1
>
> Nombre de *clusters* ayant des images se retrouvant dans 8 villes : 3
>
> Nombre de *clusters* ayant des images se retrouvant dans 6 villes : 2
>
> Nombre de *clusters* ayant des images se retrouvant dans 5 villes : 6

[14] Ce sont ces seize *clusters*-là qui ont été classifiés pour le corpus unifié.



Nombres de clusters *pour chaque niveau de diffusion spatiale des images (niveau : 35 villes à niveau : 5 villes)*

Relativement aux *clusters* ayant des images retrouvées dans 4 villes :

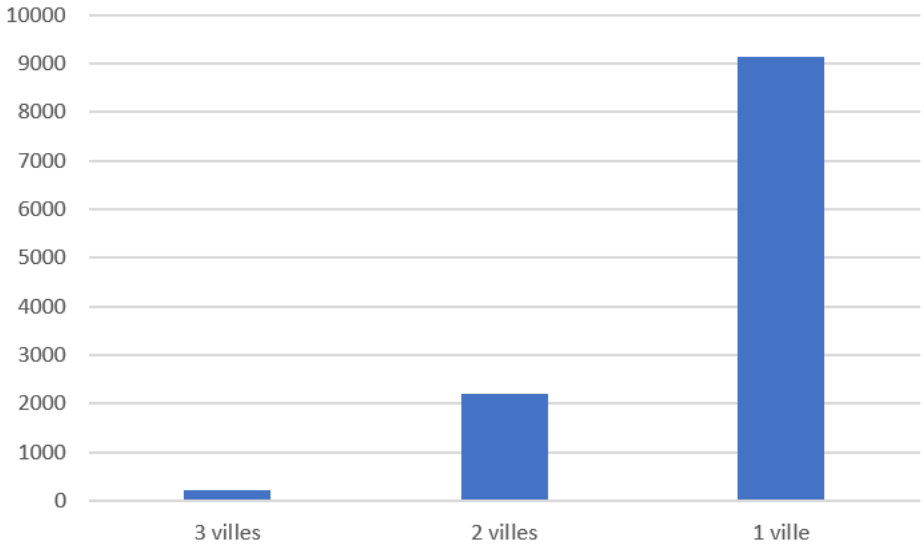
Nombre de *clusters* ayant des images se retrouvant dans 4 villes : 31

Relativement aux clusters ayant des images se retrouvant dans un nombre de villes compris entre 3 et 1 :

Nombre de *clusters* ayant des images se retrouvant dans 3 villes : 217

Nombre de *clusters* ayant des images se retrouvant dans 2 villes : 2 213

Nombre de *clusters* ayant des images se retrouvant dans 1 ville : 9 132



Nombres de clusters *pour chaque niveau de diffusion spatiale des images (niveau : 3 villes à niveau : 1 ville)*

Quelle est la moyenne des dates du corpus unifié ?
En moyenne, les images ont été publiées **en 1927**. (Obtenue grâce à des lignes de code *Python* (cf. Annexe 2).)

Quelle est la médiane des dates du corpus unifié ?
La médiane des dates au sein du corpus unifié est **l'année 1926**. (Obtenue grâce à des lignes de code *Python* (cf. Annexe 2).)

Quelles sont les revues les plus représentées au sein du corpus unifié ?
Les cinq revues les plus représentées au sein du corpus unifié sont, dans un ordre décroissant d'itérations :

- *Jugend*
- *Fliegende Blätter*
- *Meggendorfer-Blätter*
- *L'Amour de l'Art*
- *Punch*

Quelle est la médiane du nombre de villes par cluster *?*

Cette médiane est de 1.

Remarque : Dans ce corpus unifié, la plupart des clusters (9 132 sur 11 609) ont seulement des images publiées dans une seule ville, et ne témoignent donc, par définition, d'aucune circulation spatiale.

Si l'on considère seulement les *clusters* ayant des images réparties dans 4 villes au minimum, l'on doit analyser 47 *clusters*. *Faute de temps, dans le cadre de cette étude, la classification et l'analyse a été restreinte aux seize* clusters *dégagés plus haut, et non pas à ces 47* clusters*.*

Si l'on considère seulement les *clusters* ayant des images réparties dans 5 villes au minimum, l'on doit analyser et classifier seize *clusters*.

B) Classification : *clusters* d'images se retrouvant dans cinq villes au minimum

1er *cluster* : 35 villes (141 963 images) ***cluster* « incohérent »**
cluster:01383900567c7733f03e58712b3f832f531935f4e190eeb2da0d646f38fab10bb872f2b6

2e *cluster* : 19 villes (239 images) ***cluster* « stérile »**
cluster:000085003eb5ab4015e3c380232c04256512958a6a4a19d26d9d4476aa78d57bc83216cd
CODES BARRES

3e *cluster* : 17 villes (356 images) ***cluster* « stérile »**
cluster:0000c500515dfef4c36d6dc3788e03b3f1cae5cd9107a3e9b30d904ee25e60b35989d54c
QUATRIÈMES DE COUVERTURE

4e *cluster* : 15 villes (145 images) ***cluster* d'images similaires**
cluster:00004f006b5c5e7922a38df995911d3d390b3c21d2ac2dd11b9a9695f70583a0baa91100
PORTRAITS DIFFÉRENTS LES UNS DES AUTRES

5e *cluster* : 9 villes (111 images) ***cluster* d'images similaires**
cluster:00003e005d8b310b66d7bb86c9e5c9c64b061e7a9db8c1e0e0c55982adcc24381be9d7a7
INTÉRIEUR DE CATHÉDRALES

6e *cluster* : 8 villes (299 images) ***cluster* d'images similaires**
cluster:00008a0026937e52d33902dadfddae0ff6382a56338592889b2eadd4935efb20dad38930
PORTRAITS (photographies, tableaux)

7e *cluster* : 8 villes (187 images) ***cluster* d'images similaires**
cluster:00005d008a899e179849e92791e25ae9de71d7d2eecae23f1b558df80b9f2bbf3feee2f2
PORTRAITS (statues, tableaux)

8e *cluster* : 8 villes (103 images) ***cluster* d'images similaires**
cluster:000038003204951f6d337a84596c7cfc9996f8c6ee20f7ab24a2aa680da93c4e3d041aa0
VERRERIE (verres, chandeliers, vases, etc.)

9e *cluster* : 6 villes (22 images) ***cluster* d'images similaires**
cluster:00000b00e8ae4b848de27b314f3cd0e25eefd1f294f99d39da451a9058ad83e83d360fbf
DIFFÉRENTS BÂTIMENTS (villa/maison, immeuble à deux étages, etc.) ; ARCHITECTURES DE TYPE MODERNISTE (ex : l'architecte écossais Thomas S. Tait)

10e *cluster* : 6 villes (25 images) ***cluster* d'images similaires**
cluster:00000d0050c8b8a45ff8f87c29f05a7aba34b65eb73b4e2ae9816a329b9272e0803c1849
STATUES PORTRAIT (bustes)

11e *cluster* : 5 villes (63 images) ***cluster* d'images similaires**

cluster:00001f00be9faf2e91cf5157e07dd02f0b28e32e59fcf91cb1d48b40ff144c8486f86342
LUSTRES (LUMINAIRE DÉCORATIF)

12e *cluster* : 5 villes (70 images) ***cluster* d'images similaires**
cluster:000020004842ebea4de9088f3836d45ce0978d61738e1b7f295f1015f960b0c376d4fea0
DIFFÉRENTES VOITURES (*Rolls Royce*, *Peugeot*, *Cadillac*, etc.)

13e *cluster* : 5 villes (11 images) ***cluster* d'images similaires**
cluster:000006003f087659b95e98ab5792f9b046cccb81cdfc272006e844adc6031832b468fdb3
DIFFÉRENTES VOITURES (*Mercedes-Benz*, *Renault*, etc.)

14e *cluster* : 5 villes (20 images) ***cluster* d'images similaires**
cluster:00000c0034954e4ea66d219c1b37fe6eab92915a8eb9c9e6a2b824a9d165eb58b5d21e18
OBJETS LONGILIGNES (lavabo de style gothique, horloge, etc.)

15e *cluster* : 5 villes (12 images) ***cluster* d'images identiques (EXPLOITABLE)**
cluster:0000070055431aac128c87e71ec0fb7842d2ce80e32d4d2c6a1a6af2edb15735c3460ad1
Albrecht Dürer, *Melencolia I*, 1514 [gravure] ➔ **CIRCULATION POTENTIELLE**

16e *cluster* : 5 villes (9 images) ***cluster* d'images identiques (EXPLOITABLE)**
cluster:000005003e44edd200bff7f2b77a484ac8832cfcc6feb663b749680958fef3979ccc3be2
Max Beckmann, *Port de Gênes*, 1926 [tableau] ➔ **CIRCULATION POTENTIELLE**

C) Deux potentielles circulations d'images

Une analyse aussi bien quantitative (mesure du niveau de diffusion spatiale des *clusters*) que qualitative (exclusion des *clusters* incohérents, ou « stériles », ou regroupant des images seulement similaires) des *clusters* du corpus unifié, a permis d'isoler deux *clusters* susceptibles de témoigner, pour chacun d'eux, d'une circulation d'une image.

Ces deux *clusters* concernent :

D'une part, une gravure du XVI^{e} siècle, dont des photographies ont été publiées dans des revues localisées dans cinq villes différentes : Berlin, Leipzig, Munich, Vienne et Paris ; de trois différents pays : l'Allemagne, l'Autriche et la France ; entre 1921 et 1939 ;

D'autre part, un tableau du XX^{e} siècle, dont des photographies ont été publiées dans des revues localisées dans cinq villes différentes : Berlin, Munich, Copenhague, Leipzig et Paris ; de trois différents pays : l'Allemagne, le Danemark et la France ; entre 1928 et 1931.

***Melencolia I*, Albrecht Dürer, 1514**

Tel que publié dans *Der Kunstwanderer*, Berlin, Allemagne, 1927

	A	B	C	D	E	F	G	H	I	J	K
1	manifeste_u	image_url	page_url	numero_clus	titre	date	Title	Country	City	Year	normalized_date
2											
3	https://digi.u	https://digi.u	https://digi.u	cluster:0000070055431aac128c87e71ec			Das Buch fÃ¼	Germany	Berlin, Leipzi	1921.0	1921
4											
5	https://digi.u	https://digi.u	https://digi.u	cluster:0000070055431aac128c87e71ec			Kunstwart ur	Germany	Munich	1921.0	1921
6											
7	https://digi.u	https://digi.u	https://digi.u	cluster:0000070055431aac128c87e71ec			Mitteilungen	Austria	Vienna	1925.0	1925
8											
9	https://gallic	https://gallic	https://gallic	cluster:0000(	Gazette des l	janv-25	Gazette des l	France	Paris		1925/01
10											
11	https://digi.u	https://digi.u	https://digi.u	cluster:0000070055431aac128c87e71ec			Der Kunstwa	Germany	Berlin	1927.0	1927, 1928
12											
13	https://gallic	https://gallic	https://gallic	cluster:0000(	Gazette des l	juil-37	Gazette des l	France	Paris		1937/07
14											
15	https://digi.u	https://digi.u	https://digi.u	cluster:0000070055431aac128c87e71ec			Weltkunst	Germany	Berlin	1939.0	1939

Classeur *regroupant les images* clusterisées *du* cluster *« MelencoliaI_AlbrechtDurer_1514 »*
Avec notamment les informations d'ordre spatial

Port de Gênes, Max Beckmann, 1926

Tel que publié dans *Der Cicerone*, Leipzig, Allemagne, 1930

	A	B	C	D	E	F	G	H	I	J	K
1	manifeste_u	image_url	page_url	numero_clus	titre	date	Title	Country	City	Year	normalized_date
2											
3	https://digi.u	https://digi.u	https://digi.u	cluster:000005003e44edd200bff7f2b77			Die Form	Germany	Berlin	1928.0	1928
4											
5	https://digi.u	https://digi.u	https://digi.u	cluster:000005003e44edd200bff7f2b77			Der Kunstwa	Germany	Munich	1928.0	1928, 1929
6											
7	https://digi.u	https://digi.u	https://digi.u	cluster:000005003e44edd200bff7f2b77			Samleren	Denmark	Copenhagen	1928.0	1928
8											
9	https://digi.u	https://digi.u	https://digi.u	cluster:000005003e44edd200bff7f2b77			Der Cicerone	Germany	Leipzig	1930.0	1930
10											
11	https://gallic	https://gallic	https://gallic	cluster:0000(	La Renaissan	mars-31	La Renaissan	France	Paris		1931/03

Classeur *regroupant les images* clusterisées *du* cluster *« PortdeGenes_MaxBeckmann_1926 »*
Avec notamment les informations d'ordre spatial

Peut-on, dans ces deux cas, parler de circulations d'images ?

Les statistiques descriptives peuvent servir de moyen afin de confirmer ou, au contraire, d'infirmer une hypothèse. En l'occurrence, il s'agirait de confirmer l'hypothèse, ou à tout le moins l'intuition, d'une circulation d'une image, dans chacun des deux cas des *clusters* isolés plus haut.

En somme, et de manière plus générale, deux circulations – indépendantes l'une de l'autre – de deux images sont envisagées. Cette seule possibilité fait que, même sans qu'il soit question d'une *hypothèse* (dont les critères de validité scientifique les plus élémentaires ne sont peut-être pas remplis ici), toute tentative de confirmation de cette possibilité envisagée doit, afin d'avoir une force probante, viser un certain niveau de validité scientifique.

D'où l'utilité des statistiques.

Cependant, les statistiques ne revêtent un raisonnement, fondé sur elles, de la force probante scientifique que lorsqu'il est question d'un certain nombre de données. Autrement dit, un certain seuil de quantité de données doit être rempli afin de donner aux statistiques tout leur potentiel démonstratif – ce *seuil* variant selon l'objet d'étude.

Or, aussi bien en ce qui concerne le cas de la gravure d'Albrecht Dürer, que celui du tableau de Max Beckmann, il semble que les données soient trop peu nombreuses pour que les statistiques puissent jouer un rôle démonstratif valable.

Cela ne veut pas dire que toute circulation est à écarter ; loin s'en faut ! Mais, dans ces deux cas, la recherche d'éventuelles *contagions visuelles* devra, semble-t-il, se placer sur le terrain de l'analyse historique ; et non pas sur celui de l'analyse statistique en tant que telle. Dans cette optique, il faudra peut-être collecter d'autres informations, notamment relatives aux contextes historiques des images en question, afin de tenter d'appuyer l'idée de circulations d'images, dont le moteur des flux concernés serait la *contagion visuelle*.

En définitive, relativement à ces deux cas-là, nous pouvons seulement constater la présence d'images identiques à différents endroits sur des années différentes. Et ce, même s'il serait tentant de relier ces différents *points*, ayant chacun des *coordonnées spatiales* et *temporelles*, avec des flux (illustrés par des flèches).

À titre d'exemple :

On sait qu'une image de la gravure d'Albrecht Dürer a été publiée à Paris en 1925 ; à Berlin en 1927 ; à Paris en 1937 ; à Berlin en 1939.

Un raisonnement *hâtif* aboutirait à en conclure que cette image de la gravure a circulé en suivant le sens Paris-Berlin, puis le sens Berlin-Paris, puis, encore une fois, le sens Paris-Berlin.

Nous ne pouvons pas prouver cela, en tout cas pas *via* une analyse statistique.

La seule conclusion que l'on puisse tirer de cet exemple est la présence à différents endroits, sur différentes années, des images de cette gravure d'Albrecht Dürer.

Cinquième partie – Discussion et Perspectives

Méthodologie d'analyse proposée – jeu de données plus adapté à la recherche de circulations d'images

Le jeu de données ayant servi de base à cette étude est composé d'une part conséquente de *clusters* inexploitables. C'est d'ailleurs pourquoi le travail de classification partielle de ces *clusters* revêt une telle importance. Idéalement, la recherche d'éventuelles circulations d'images, de contagions visuelles, devrait se baser sur des données purgées, autant que faire se peut – et ce de quelque manière que ce soit (bien sûr, une automatisation d'une telle tâche serait ici bienvenue), de tels *clusters* inexploitables, et notamment, dans un premier temps, des *clusters* « incohérents ».

Cela étant dit, il est intéressant d'envisager une méthodologie d'analyse de telles données, *nettoyées*, et donc plus adaptées à la recherche de circulations d'images.

Une telle méthodologie pourrait s'axer sur deux lignes, complémentaires l'une de l'autre :

- L'hypothèse d'une corrélation : *cluster* volumineux[15] *égal* plus de chance de découvrir des circulations d'images par contagions visuelles ;
- L'hypothèse d'une corrélation : *cluster* avec forte diffusion spatiale[16] *égal* plus de chance de découvrir des circulations d'images par contagions visuelles.

S'agissant de la première corrélation, nous avons pu observer que cette corrélation n'était pas forcément vérifiable. Mais cela principalement en raison des *clusters* inexploitables, et parmi ceux-là notamment les *clusters* « incohérents », qui, souvent très volumineux, ne démontraient pas pour autant la présence, en leur sein, de signes de circulations.

C'est pourquoi un jeu de données purgé de tels *clusters* serait le bienvenu est permettrait de vérifier précisément cette première corrélation. En effet, dire que plus un *cluster* serait volumineux, plus il aurait de chance de présenter les signes d'une ou de circulations apparaît comme logique dans l'absolu ; mais il resterait à savoir dans quelle mesure s'inscrirait cette affirmation.

À cet effet, il s'agirait de repérer les clusters les plus volumineux et s'intéresser, dans un premier temps, notamment à ceux dans lesquels les années des dates de publication des images seraient les plus *étalées*. En effet, le projet *Visual Contagions* concerne l'ère pré-Internet ; dès lors, il est difficilement envisageable d'avoir des contagions visuelles s'inscrivant dans un temps court. Les circulations d'images par contagion visuelle, en la matière, seront le plus souvent, semble-t-il, affaire de *temps long* – et ce, notamment si la recherche s'axe sur la question de la circulation mondiale des styles. Mais, rien n'est à exclure.

S'agissant de la seconde corrélation, elle est plus évidente que la première.

En effet, constater que des images se retrouvent à beaucoup d'endroits c'est constater que, d'une manière ou d'une autre, il y a eu circulation. Néanmoins, il reste à savoir si cette circulation s'est faite par *contagion visuelle*.

[15] *Cluster* volumineux = *cluster* avec beaucoup d'images
[16] *Cluster* avec forte diffusion spatiale = pas forcément beaucoup d'images (loin s'en faut !) mais des images présentes dans de nombreuses villes différentes (ou pays, etc.)

À cet effet, il s'agira de vérifier si la circulation répond aux critères d'une *contagion visuelle* ; en définissant précisément l'étendue de la notion de *contagion visuelle*. Et, concrètement, d'un point de vue méthodologique, un jeu de données plus adapté permettrait, à ce niveau, que soit faite une distinction selon que la recherche de circulations d'images concerne : la circulation mondiale des styles sur deux siècles ; la diffusion par l'affiche des vocabulaires visuels radicaux ; ou l'expansion mondiale de nouvelles images de la femme dans la presse illustrée depuis 1945.

En effet, le premier thème ne nécessiterait peut-être pas de filtrer les revues ; même si une attention particulière pourrait être accordée au sein des *clusters* aux revues d'art – mais ce serait peut-être risquer de *passer à côté* de circulations : un artiste ne puise pas forcément son inspiration (notamment stylistique) de revues académiques au sens large (peut-être qu'en la matière cela dépendra des époques).

Le second thème orienterait plus la recherche vers les revues à audience large, voire très large ou, au contraire, vers des revues à audience restreinte mais très spécialisée : autour de partis/mouvements politiques, de syndicats, etc.

Le dernier thème, de par son intitulé – « […] dans la presse illustrée depuis 1945 » – nécessiterait un filtre des revues.

Piste de réflexion

Exponentielle vs *Affine à coefficient directeur élevé*

Dans le cadre d'une recherche de contagions visuelles, découvrir une contagion qui se serait propagée de manière exponentielle serait très intéressant, d'autant plus si elle concernait l'époque pré-Internet[17].

Quel serait le « point de départ » d'une contagion visuelle exponentielle ?

Le point de départ d'une contagion exponentielle serait, semble-t-il, plus susceptible d'être décelé dans une revue à faible audience.

Pourquoi ?

En raison du fait qu'une revue à forte audience, voire très forte audience – comme *Le Monde* ou *The Times* – a plus de chance d'être le point de départ non pas d'une contagion exponentielle mais d'une contagion qui suivrait une courbe (plus ou moins) linéaire, d'une fonction affine ; mais à coefficient directeur élevé, voire très élevé. Une telle contagion serait certes *puissante*, car elle reposerait avant tout sur l'audience élevée d'une telle revue, qui expliquerait le coefficient directeur élevé, mais elle serait peut-être moins *impressionnante* qu'une contagion exponentielle.

En effet, une contagion visuelle exponentielle serait *impressionnante*, dans le sens où la différence des quantités de contagions[18] entre un instant t et un instant $t + 1$ (des instants qui ne seraient pas tellement éloignés l'un de l'autre), serait très élevée.

[17] Les diffusions exponentielles sur l'Internet sont facilement envisageables.
[18] 3 contagions un jour ; 7 contagions un autre jour, etc.

À l'observation d'une telle courbe de contagion visuelle, il faudra comprendre que l'évolution des cas de *contagions visuelles* aura toujours suivi une courbe exponentielle, que ce soit au début dans le cadre de l'audience très faible d'une revue *quelconque*, ou en plein *pic* dans un cadre de diffusion de l'image concernée devenu européen voire intercontinental.

Conclusion

Synthèse

Dans ce rapport statistique, nous avons pu mettre en évidence l'existence d'une variété de *clusters*, que nous avons pu regrouper au sein de différentes catégories. Ce faisant, ce rapport peut s'apparenter à un support de travail susceptible de permettre un gain d'efficacité dans le cadre de la recherche de circulations d'images par contagion visuelle.

Dans cette optique, un premier pan de ce rapport offre une description statistique permettant de cerner les limites de *Visual Contagions Explore* (*VCE*), et de pouvoir peut-être, à l'avenir, modifier la plateforme en vue de dépasser ces limites. Nous pensons par exemple au fait que *VCE* « rejette » un certain pourcentage de manifeste. Mais encore à la question de la segmentation des images des manifestes.

Un second pan de ce rapport, vise à préparer une étude sur un jeu de données plus adapté à l'étude de circulations d'images, en opérant une classification partielle des *clusters*, qui pourrait être assimilée à un début de *nettoyage des données*. À tout le moins, le travail de classification des *clusters* opéré sur le corpus unifié (et même celui opéré sur l'ensemble de quatorze corpus) peut éventuellement servir de *clé de voûte* à l'établissement d'une stratégie de *nettoyage des données*.

Par un ensemble de précisions contextuelles, ce rapport statistique apparaît aussi comme une ressource afin de comprendre le fonctionnement de *VCE*, et par conséquent l'utiliser de la manière la plus idoine pour préparer une étude statistique visant à découvrir des circulations d'images.

Ce travail statistique est ponctué par la mise en exergue du cas de deux potentielles circulations, et c'est en gardant une approche statistique de la question que nous avons dû nous résigner à ne pas pouvoir nous prononcer sur l'existence ou pas en l'espèce de circulations d'images. En effet, seuls des *clusters* (exploitables, bien entendu) composés d'un nombre conséquent d'images pourraient permettre de tirer des conclusions appuyées par la force probante scientifique des statistiques. Il reste que la potentialité de circulations, au sein des deux *clusters* isolés, est réelle.

Analyse historique

Le défaut d'observation des clusters pertinent pourrait en outre tenir d'une explication historique. De fait, la période proposée à l'analyse, entre 1920 et 1939, fait partie selon l'historien Eric John Hobsbawm de l'Âge des catastrophes qu'il situe entre 1914 et 1945[19]. Périodicité tristement marquée par un déchirement interne causé par la Première Guerre Mondiale et ses suites, la période d'entre-deux-guerres débutant en 1918, soit à la veille de notre corpus, est marquée par un désordre profond et durable. Les Empires ne sont plus, des idéologies se cristallisent[20], des conflits locaux voient leur désastreuse apogée. Tout en Europe crie à la brutalité la plus perverse. La misère s'avoue humaine, sociale, politique, militaire mais aussi industrielle.

Le corpus de notre étude met en lumière les relations – ou non relations – franco-allemandes. De fait, des suites du Traité de Versailles de 1919 vécu comme un véritable diktat par une partie du peuple allemand, la France reprend le contrôle sur l'Alsace-Moselle. D'autres tensions fortes émergent dans les années 1920, notamment pour des raisons de répartitions financières, et à cela vient s'ajouter l'Accord d'assistance militaire franco-belge de 1920, qui promeut l'occupation de la Ruhr pour remédier aux retards de payements des indemnités de guerre. Cet enchainement d'événements fait naitre un profond sentiment anti-français ainsi qu'un esprit de revanchisme[21]. Ces sentiments perdurent, se nationalisent et se concrétisent en l'ascension de la Seconde Guerre Mondiale.

Sur le plan artistique, un constat similaire peut être établi. L'entre-deux-guerres s'avoue être une des périodes les plus fécondes pour l'art allemand promouvant une innovation iconographique et stylistique notamment sous l'égide de la *Neue Sachlichkeit,* du Bauhaus ou encore du rayonnement international du Dada[22]. Du côté français, l'impressionnisme s'épuise lentement. Ainsi, bien que l'art s'internationalise, les différends franco-allemands cristallisent leur langage artistique.

[19] HOBSBAWN Eric, *Age of Extremes – The short Twentieth Century 1914-1991,* Londres, Abacus, 1994.

[20] THIESSE Anne-Marie, *La Création des identités nationales. Europe, 18e-20e siècle*, Paris, Editions du Seuil, 1999.

[21] JOLY Bertrand, « La France et la Revanche », *Revue d'histoire moderne et contemporaine*, Vol. 46 (2), 1999, pp. 325-347.

[22] GISPERT Marie, *« L'Allemagne n'a pas de peintres ». Diffusion et réception de l'art allemand moderne en France durant l'entre-deux guerres, 1918-1939*, Thèse de doctorat en histoire de l'art préparée sous la direction de Philippe Dagen, université Paris 1-Panthéon-Sorbonne, soutenue le 9 décembre 2006, Trajectoires [En ligne], 1 | 2007, mis en ligne le 16 décembre 2009, consulté le 18 mai 2021. URL : http://journals.openedition.org/trajectoires/95 ; DOI : https://doi.org/10.4000/trajectoires.95.

Ouverture

Tout d'abord, nous invitons d'ores et déjà les personnes souhaitant poursuivre l'étude, à s'intéresser dans le cadre du corpus unifié, aux *clusters* ayant des images dans au minimum quatre villes (et non pas cinq). Cela ajouterait 31 *clusters* à étudier : d'autres potentielles circulations – en sus des deux qui ont pu être isolées dans notre étude – pourraient peut-être apparaître.

Ensuite, parmi les *clusters* classifiés comme inexploitables, certains pourraient revêtir d'un intérêt selon les questions envisagées. Nous pensons particulièrement au cas des publicités qui pourrait, par exemple, intéresser des questions d'histoire de l'économie.

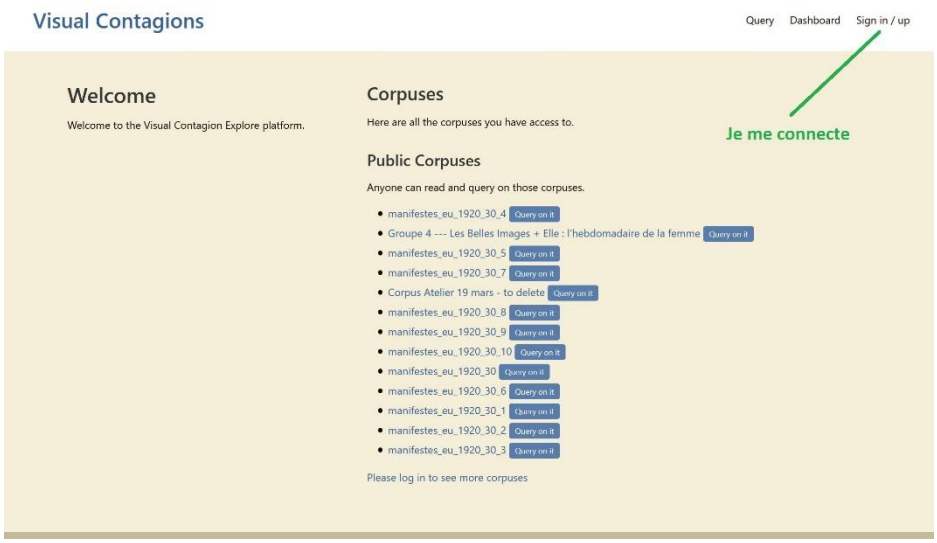
Enfin, une approche pluridisciplinaire pourrait intégrer l'épidémiologie. En effet, s'agissant de *contagions* visuelles, il pourrait être intéressant d'utiliser les modèles compartimentaux en épidémiologie. Ces modèles, à visée prédictive, permettent de cibler les groupes susceptibles d'être contaminés. Or, malgré *leur visée prédictive*, ces modèles auraient un intérêt à être envisagés en histoire. En effet, grâce à une expertise en histoire de l'art par exemple, il pourrait être intéressant de déterminer des groupes susceptibles d'être l'objet de *contagions visuelles* entre eux (ex : des revues appartenant s'inscrivant dans le même courant artistique), et dès lors créer des modèles ou même des *proto*-modèles de contagions visuelles. Des modèles qui, et c'est là où réside l'avantage d'étudier des phénomènes historiques, pourraient ensuite être vérifiés en étudiant par exemple des archives historiques. Concrètement, ces modèles pourraient être un moyen d'orienter la recherche historique, voire de rationaliser un choix d'étude historique : telle période et tels espaces seraient privilégiés car répondant favorablement aux modèles statistiques de contagion visuelle.

Annexes

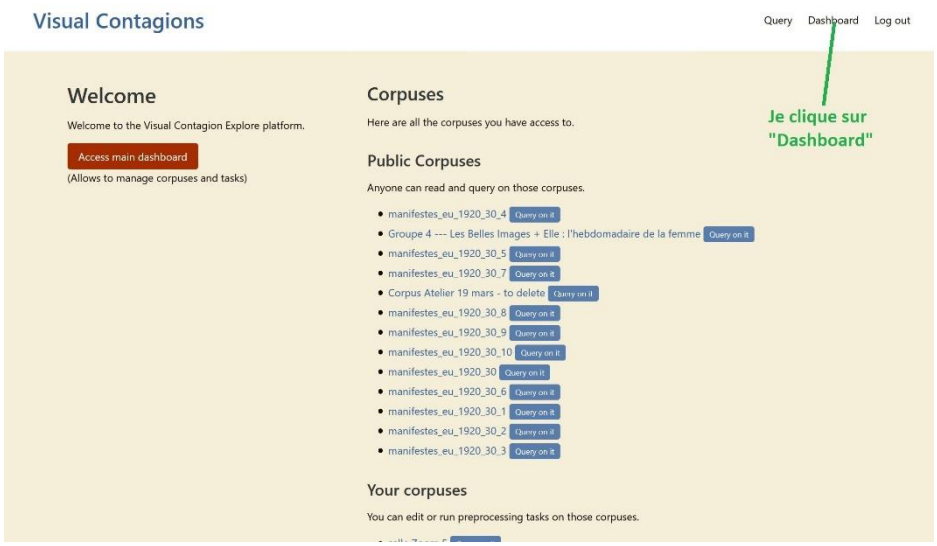
Annexe 1 – Procédure d'utilisation de la plateforme (Fournir *VCE* en URLs et obtenir des *clusters* ; « nettoyer » les corpus)

Fournir VCE *en URLs de manifestes et obtenir des* clusters *d'images*

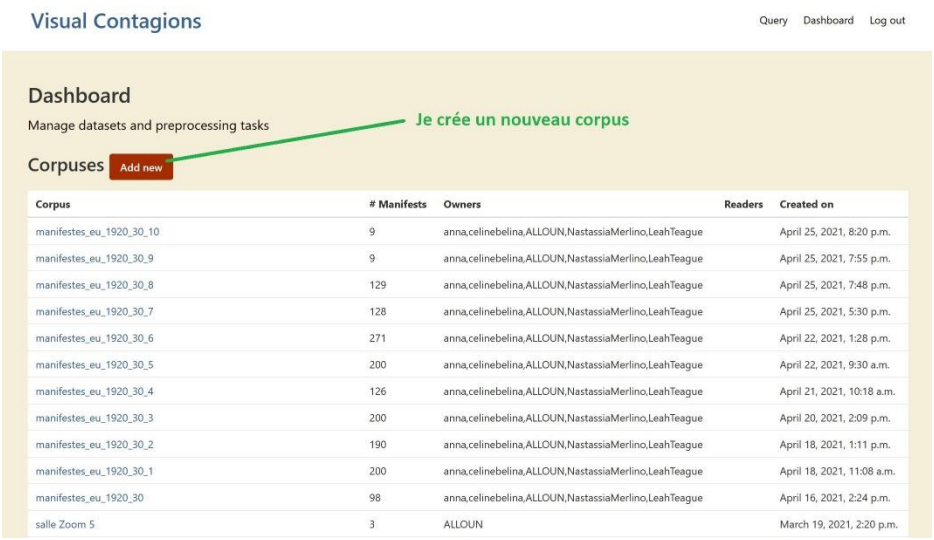
Avant de pouvoir récupérer des *clusters*, il s'est agi de fournir la plateforme *VCE* en URLs de manifestes. Voici des captures d'écran, permettant de comprendre – grâce à une narration à la première personne du singulier – la démarche à suivre.



Connexion

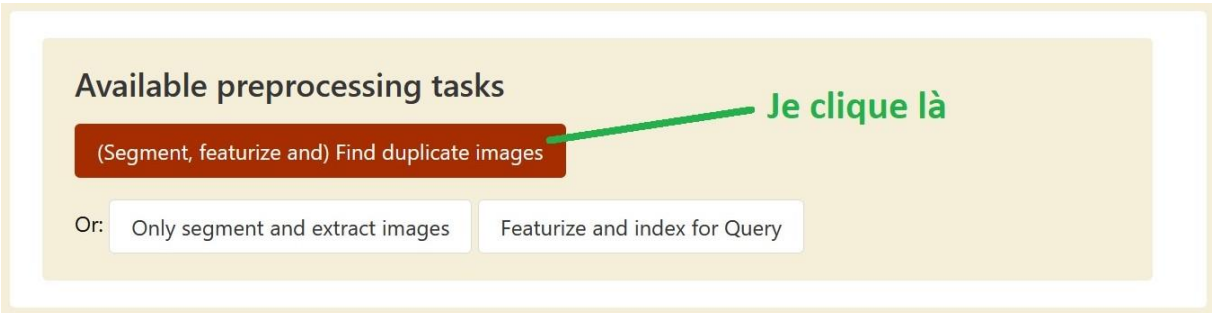


Accéder au « dashboard »

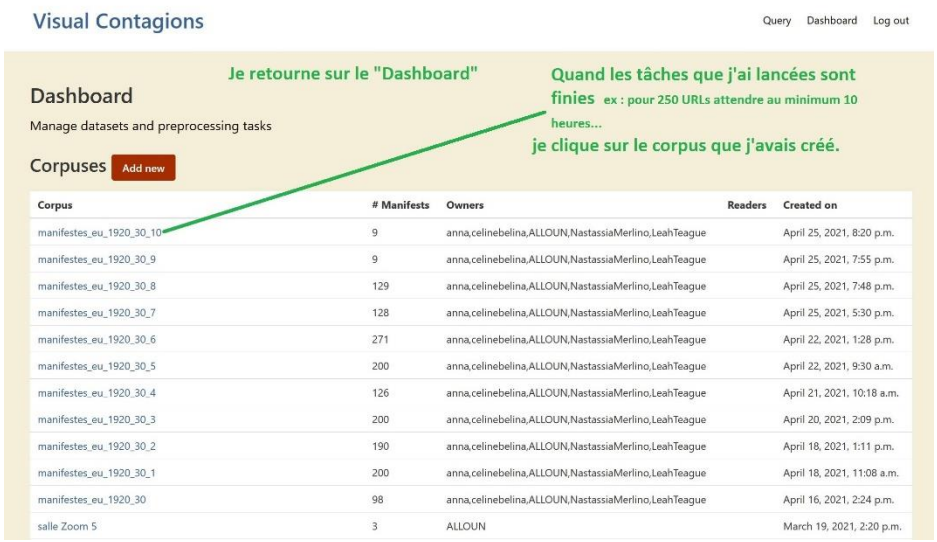


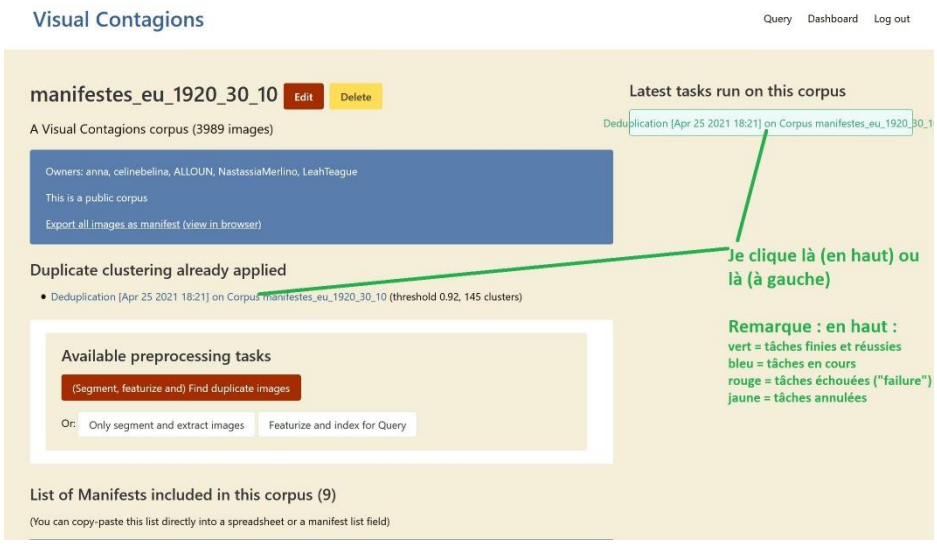
Création d'un nouveau corpus

Copie-collage des URLs

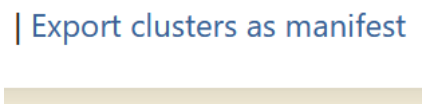


Lancement des tâches ayant notamment pour but d'obtenir des clusters





Accès aux clusters

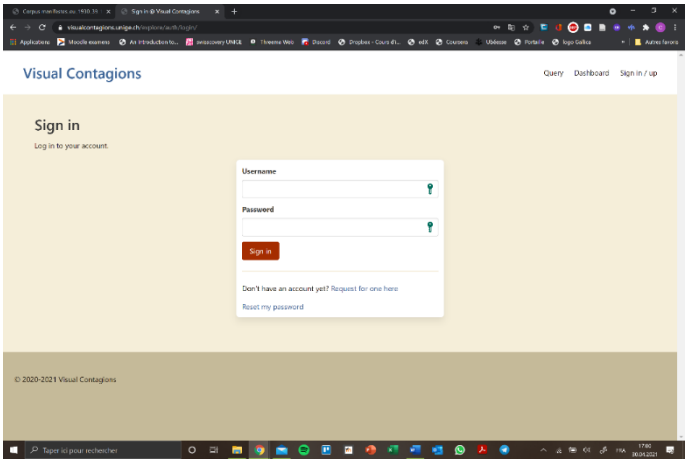


Exportation des clusters via *téléchargement (là où il faut cliquer)*

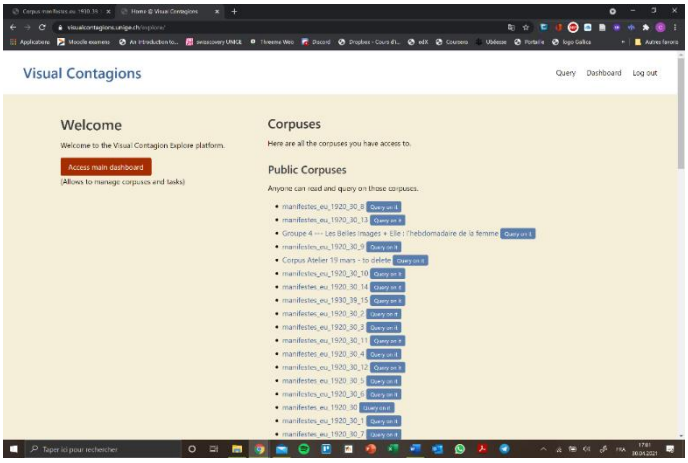
« Nettoyer » les corpus afin de ne garder que des manifestes avec images

Dans l'optique de la fusion des quatorze corpus en un seul : le corpus unifié, il s'est agi préalablement de « nettoyer » les corpus. Autrement dit, ont été retirées de chacun des corpus les URLs relatives à des manifestes ne contenant pas d'images. Voici des captures d'écran, permettant de comprendre, étape par étape, la démarche à suivre en la matière.

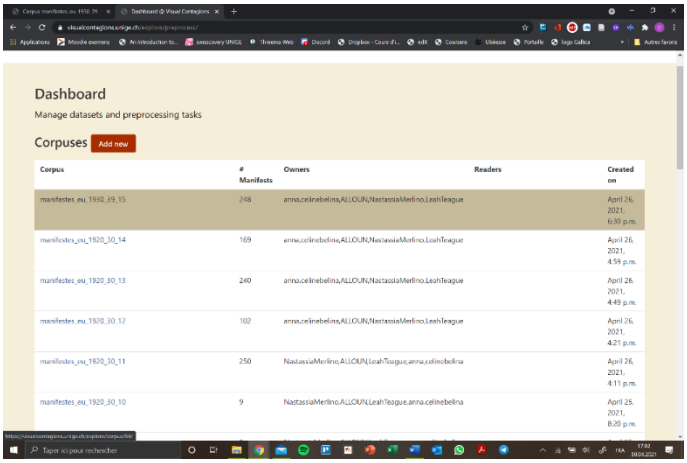
1. Aller sur *Visual Contagions Explore* https://visualcontagions.unige.ch/explore/ et s'enregistrer avec son login.



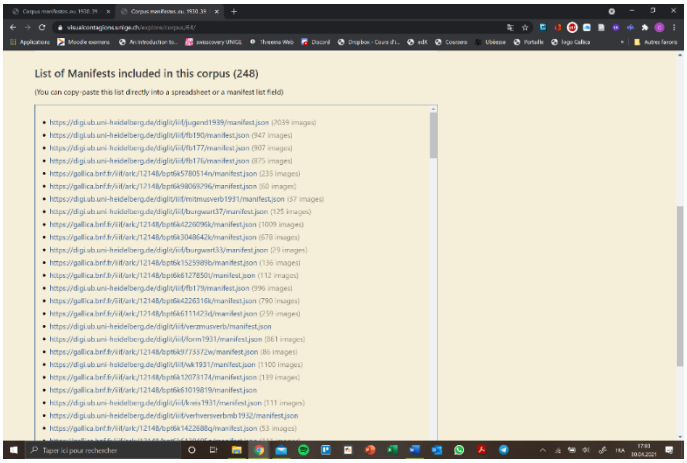
2. Appuyer sur « Access main dashboard » (bouton rouge) ou « Dashboard ». (Dans notre cas le « Dashboard » principal (*main*) était unique ; d'où la possibilité de cliquer sur l'un ou l'autre bouton – à l'avenir, il en ira peut-être autrement.)



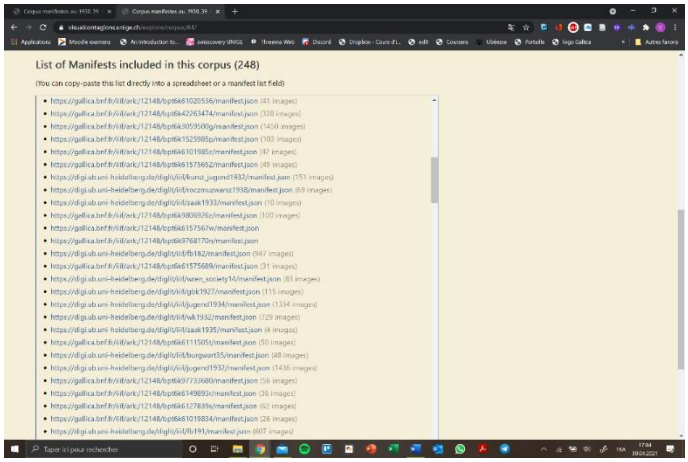
3. Dans les « Corpuses », appuyer sur un des manifestes_eu_19xx_xx_xx



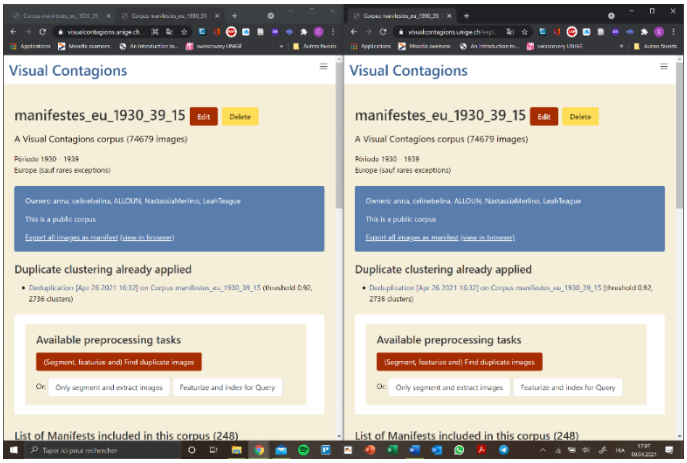
4. Descendre sur la même page jusqu'à « List of Manifests included in this corpus »



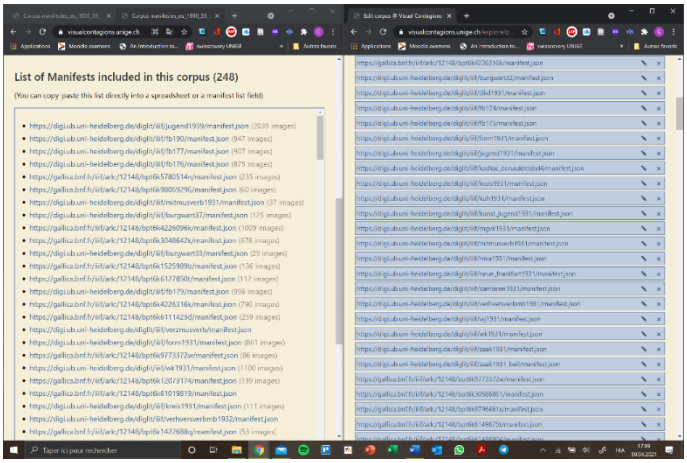
5. Repérer les corpus sans images, c'est-à-dire ceux qui n'ont pas de parenthèses grisées à côté de l'URL avec (xx images)



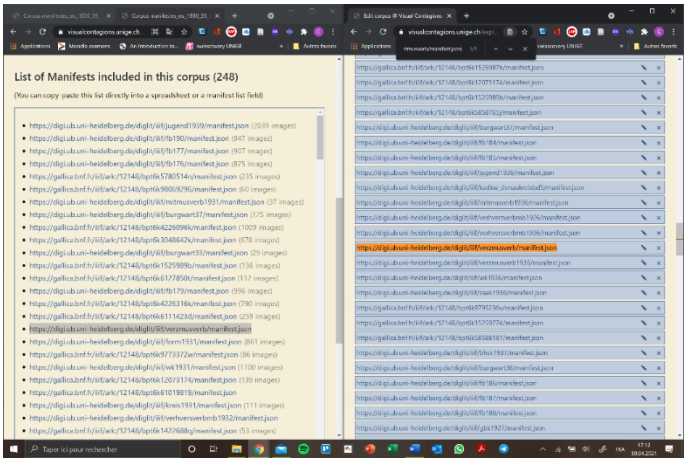
6. Copier le lien de cette page et en ouvrir une fenêtre identique à côté



7. Sur la fenêtre de droite cliquer sur le bouton rouge « Edit » et descendre sur les deux pages jusqu'à visionnement des corpus



8. Regarder sur la page de droite les manifestes sans images et les supprimer à l'aide de la petite croix sur la page de gauche. (Possibilité d'utiliser « Ctrl + F » afin d'être sûr d'avoir le même manuscrit.)



Annexe 2 – *Jupyter Notebooks* (en langage *Python*)

Jupyter Notebooks avec des programmes informatiques contenant du code et du texte

Rédigés par le groupe d'étudiants n° 6 (*GT 6*)

Dans le cadre du cours transversal : Comprendre le Numérique

Université de Genève

Creation_CSV_images_clusterisees

May 25, 2021

1 Installation et importation de librairies Python

[ ]:
```
!pip install requests
!pip install pandas
```

[ ]:
```
import os
import json
import requests
import pandas as pd
```

2 Chemin du fichier JSON

[ ]:
```
os.getcwd()
```

[ ]:
```
os.listdir(os.getcwd())
```

[ ]:
```
os.listdir('..')
```

3 Charger le contenu du fichier JSON

[ ]:
```
clusters_file = "manifest.json"
with open(clusters_file) as f:
    data = f.read()
clusters = json.loads(data)
```

4 Vérification

Vérifiez que les lignes de code suivantes permettent bien d'accéder aux informations souhaitées.

[ ]:
```
clusters["resources"][0]["resource"][2]["chars"] # lien manifeste
```

[ ]:
```
clusters["resources"][0]["resource"][1]["chars"] # numéro cluster
```

[ ]:
```
clusters["resources"][0]["resource"][3]["@id"] # url image
```

[ ]:
```
clusters["resources"][0]["resource"][3]["full"]["@id"] # url page
```

Si ce n'est pas le cas, modifiez-les en accédant aux informations d'une image, grâce à la ligne de code suivante. ([0] désigne la première image ; changez de nombre afin d'accéder aux informations d'une autre image.)

[ ]:
```
print(json.dumps(clusters["resources"][0], sort_keys=True, indent=2))
```

5 Création d'une liste contenant les informations de toutes les images du fichier JSON

[ ]:
```
images = []
for i, resource in enumerate(clusters["resources"]):
    image = {}
    try:
        image['manifeste_url'] = resource["resource"][2]["chars"]
        image['image_url'] = resource["resource"][3]["@id"]
        image['page_url'] = resource["resource"][3]["full"]["@id"]
        image['numero_cluster'] = resource["resource"][1]["chars"]
        images.append(image)
    except:
        print(i, "ème images n'a pas pu être parsée")
```

Vous pouvez afficher le nombre d'images de cette liste.

[ ]:
```
len(images)
```

6 Récupérer les métadonnées des images

Affichez le lien du manifeste de la première image, par exemple. (Rappel: [0] désigne la première image ; changez de nombre si vous voulez une autre image.)

[ ]:
```
images[0]["manifeste_url"]
```

Chargez son contenu dans une variable.

[ ]:
```
manifeste_brut = requests.get(images[0]["manifeste_url"]).content
manifeste = json.loads(manifeste_brut, encoding="utf8")
```

6.1 Vérification

Vérifiez que les lignes de code suivantes permettent bien d'accéder aux informations : titre ; et date.

[ ]:
```
manifeste["metadata"][5]["value"] # titre
```

[ ]:
```
manifeste["metadata"][6]["value"] # date
```

Si ce n'est pas le cas, modifiez-les en accédant aux informations (notamment les métadonnées) de l'image que vous avez choisie dans la variable. Accédez à ces informations en affichant le contenu du manifeste, grâce à la ligne de code suivante. Remarque : Si vous ne trouvez pas les informations relatives au titre et à la date, ce n'est pas un problème en soi ; cela voudra simplement dire que dans votre fichier CSV il n'y aura pas ces deux informations-là.

[ ]:
```
print(json.dumps(manifeste, sort_keys=True, indent=2))
```

7 Pour l'ensemble des images du fichier JSON

[ ]:
```
images_all = images[:]
```

Vous pouvez afficher, pour chaque image de votre liste, les informations respectives des images.

[ ]:
```
len(images_all)
images_all
```

8 Le fichier CSV avec les images clusterisées

[ ]:
```
for image in images_all:
    image_url = image['manifeste_url']
    print(image["image_url"])
    try:
        image_data = requests.get(image_url).content
        image_manifest = json.loads(image_data)
        for meta in image_manifest["metadata"]:
            if "date" not in image.keys() and meta["label"] in ["Date", "date"]:
                image["date"] = meta["value"].strip()
            if "titre" not in image.keys() and meta["label"] in ["Title",␣
 ↪"title"]:
                image["titre"] = meta["value"].strip()
    except:
        print(image_url, ": doesn't work")
print("Terminé !")
```

[ ]:
```
df = pd.DataFrame(images_all)
```

[ ]:
```
df
```

8.1 (Si vous souhaitez modifier l'agencement des colonnes)

[ ]:
```
df.columns.values
```

Par exemple : mettre la colonne "numero_cluster" en premier…

[ ]:
```
df = df[['numero_cluster', 'manifeste_url', 'image_url', 'page_url', 'titre',␣
 ↪'date']]
print(df.columns.values)
```

8.2 Enregistrer le fichier CSV

[ ]:
```
nom_fichier_sauvegarde = 'NOM_DE_VOTRE_FICHIER.csv'
df.to_csv(nom_fichier_sauvegarde, sep='\t', index=False)
```

Fusion fichier pour la mesure de diffusion spatiale

May 25, 2021

[ ]:
```
import os
import json
import requests
import pandas as pd
```

0.1 Réaliser la fusion du classeur (sous format CSV) et de notre fichier CSV d'images clusterisées

[ ]:
```
backup = r"C:\Users\exemple\Desktop\dossierexemple\classeur.csv"
```

[ ]:
```
tableur = pd.read_csv(backup, delimiter=';')
```

[ ]:
```
tableur.head()
```

[ ]:
```
tableur2 = tableur[["Media URL","Title", "Country", "City", "Year",␣
↪"normalized_date"]]
tableur2.head()
```

[ ]:
```
pathTab3 = r"C:
↪\Users\allou_j6awgjz\Desktop\15MAI2021\images_clusterisees_corpusMERGED_11609.
↪csv"
tableur3 = pd.read_csv(pathTab3, delimiter='\t')
tableur3.head()
```

[ ]:
```
fusion = pd.merge(tableur3, tableur2, how="left", left_on="manifeste_url",␣
↪right_on="Media URL")
```

[ ]:
```
fusion
```

[ ]:
```
nom_fichier_sauvegarde = 'fusion_clusters_lieux_corpusMERGED_11609.csv' # là␣
↪vous pouvez nommer votre fichier CSV fusionné
fusion.to_csv(nom_fichier_sauvegarde, sep='\t', index=False)
```

[ ]:
```
tableau_croise = pd.pivot_table(fusion,index=['numero_cluster'],aggfunc={'City':
↪pd.Series.nunique,'image_url':pd.Series.nunique}).reset_index('numero_cluster')
tableau_croise
```

0.2 Avoir dans un fichier CSV tous les clusters avec pour chacun leur niveau de diffusion spatiale

[ ]:
```
nom_fichier_sauvegarde = 'tableau_croise_clusters_lieux_corpusMERGED_11609.csv'
tableau_croise.to_csv(nom_fichier_sauvegarde, sep='\t', index=False)
```

[ ]:
```
decroissant = tableau_croise.sort_values('City', ascending=False)
decroissant
```

0.3 Avoir dans un fichier CSV tous les clusters avec pour chacun leur niveau de diffusion spatiale, classés dans un ordre décroissant

[ ]:
```
nom_fichier_sauvegarde =␣
↪'decroissant_tableau_croise_clusters_lieux_corpusMERGED_11609.csv'
decroissant.to_csv(nom_fichier_sauvegarde, sep='\t', index=False)
```

Manipulation_donnees

May 25, 2021

[ ]:
```
!pip install pandas
```

[ ]:
```
import os  # interagir avec les fichiers de son ordinateur
import pandas as pd  # gérer les csv
from collections import Counter  # dénombrer des éléments facilement
import numpy as np
```

[ ]:
```
os.getcwd()
```

[ ]:
```
os.listdir()
```

[ ]:
```
csv =  ("data/NotreFichier.csv") # notre fichier CSV fusionné avec les␣
 ↪informations d'ordre spatial, dans un dossier data que nous avons créé
```

[ ]:
```
tableur = pd.read_csv(csv, delimiter='\t')
```

[ ]:
```
tableur.head()
```

[ ]:
```
tableur
```

1 Manipulation de données

[ ]:
```
clusters = {}

for cluster in tableur['numero_cluster']:
    if cluster not in clusters:
        clusters[cluster] = 1
    else:
        clusters[cluster] += 1

clusters_tries = sorted(clusters.items(), key=lambda x: x[1], reverse=True)
print(clusters_tries)
```

[ ]:
```

```

```
old_values = ['cluster:
 ↪c295f3d143aa8e898ece10fba6602cc339b1c075b645073d8258465f8aeecdb2', 'cluster:
 ↪8eb830004e9ab5af5ace831dd5860c603df9214cc1e868b4dd9aadcba2ac5622', 'cluster:
 ↪cf49fbd9a5c07bcf271f216501ca2b362d60287c8fed4a9837f70b0a778f1ab4'] # trois␣
 ↪exemples de clusters intitulés comme tels par Visual Contagions Explorer
new_values = ["Cluster_1", "Cluster_2", "Cluster_3"] # à renommer en fonction␣
 ↪des résultats obtenus avec la cellule précédente

replacement_dict = {k:v for k,v in zip(old_values, new_values)}  # dictionnaire␣
 ↪avec comme clés les anciens noms de clusters et en valeur les nouveaux noms

tableur = tableur.replace(replacement_dict)  # remplacer les valeurs dans le␣
 ↪tableur
```

[ ]:
```
tableur['numero_cluster']
```

[ ]:
```
clusters = {}

for cluster in tableur['numero_cluster']:
    if cluster not in clusters:
        clusters[cluster] = 1
    else:
        clusters[cluster] += 1

clusters_tries = sorted(clusters.items(), key=lambda x: x[1], reverse=True)
print(clusters_tries)
```

[ ]:
```
Counter(tableur['numero_cluster'])
```

[ ]:
```
Counter(tableur['numero_cluster'].sort_values(ascending=True).tolist()).
 ↪most_common() # pour trier le résultat par ordre décroissant
```

[ ]:
```
Counter(tableur['numero_cluster'].sort_values(ascending=True).tolist()).
 ↪most_common(x) # pour afficher les x clusters avec le plus d'images ;␣
 ↪remplacer x par un nombre entier naturel, exemple : 3
```

[ ]:
```
dic_cluster = dict(Counter(tableur['numero_cluster'])) # Transforme la liste de␣
 ↪tuple [('Cluster_1', 160), ('Cluster_2', 71), ...] en dictionnaire :␣
 ↪{'Cluster_2': 71, 'Cluster_1': 160, ...}
print(dic_cluster)
```

[ ]:
```
pd.Series([dic_cluster[k] for k in dic_cluster]).mean() # moyenne images par␣
 ↪cluster
```

[ ]:
```
pd.Series([dic_cluster[k] for k in dic_cluster]).median() # médiane images par␣
 ↪cluster
```

Si vous souhaitez calculer la moyenne et ou la médiane d'une autre colonne du CSV, remplacez

dic_cluster par un dictionnaire des valeurs de votre colonne.
Exemple : la moyenne et la médiane du titre de revue.

[ ]:
```
dic_type = dict(Counter(tableur['titre']))
print(dic_type)
print(pd.Series([dic_type[k] for k in dic_type]).mean())
print(pd.Series([dic_type[k] for k in dic_type]).median())
```

[ ]:
```
Counter(tableur['titre'].sort_values(ascending=True).tolist()).most_common(x) #␣
↪toujours dans l'exemple de la cellule précédente, permet d'afficher les x␣
↪supports les plus utilisés ; remplacer x par un nombre entier naturel, exemple␣
↪: 4
```

[ ]:
```
print(tableur['Year'].mean()) # moyenne des dates, avec décimales
print(tableur['Year'].median()) # médiane des dates, avec décimales
```

[ ]:
```
int(tableur['Year'].mean()) # moyenne des dates, sans décimales
```

[ ]:
```
int(tableur['Year'].median()) # médiane des dates, sans décimales
```

[ ]:
```
print(min(tableur['Year'])) # année la plus ancienne
print(max(tableur['Year'])) # année la plus récente
```

Moyenne et médiane pour un cluster

Exemple : Cluster_1

[ ]:
```
mask_cluster_cluster1 = tableur['numero_cluster'] == 'Cluster_1'
print(mask_cluster_cluster1)
```

[ ]:
```
cluster_cluster1 = tableur[mask_cluster_cluster1]
cluster_cluster1
```

[ ]:
```
len(cluster_cluster1)
```

S'agissant des années…

[ ]:
```
print(int(cluster_cluster1['Year'].mean()))
```

[ ]:
```
print(int(cluster_cluster1['Year'].median()))
```

S'agissant du titre…

[ ]:
```
dic_type = dict(Counter(cluster_cluster1['titre']))
print(dic_type)
print(int(pd.Series([dic_type[k] for k in dic_type]).mean()))
print(int(pd.Series([dic_type[k] for k in dic_type]).median()))
```