

Corpus Biographes version 1.0 : construire, analyser, cartographier

**26 janvier 2017
FMSH**

**Nadège Lechevrel
Philippe Gambette**

Plan

- Présentation du corpus Biographes version 1.0
- Construire le corpus (Google Drive, PHP+MySQL)
- Analyser le corpus (TXM, TreeCloud, CFP)
- Cartographier le corpus (Palladio, auteurs citants/cités)
- Archiver le corpus (HumaNum)

Présentation du corpus Biographes 1.0

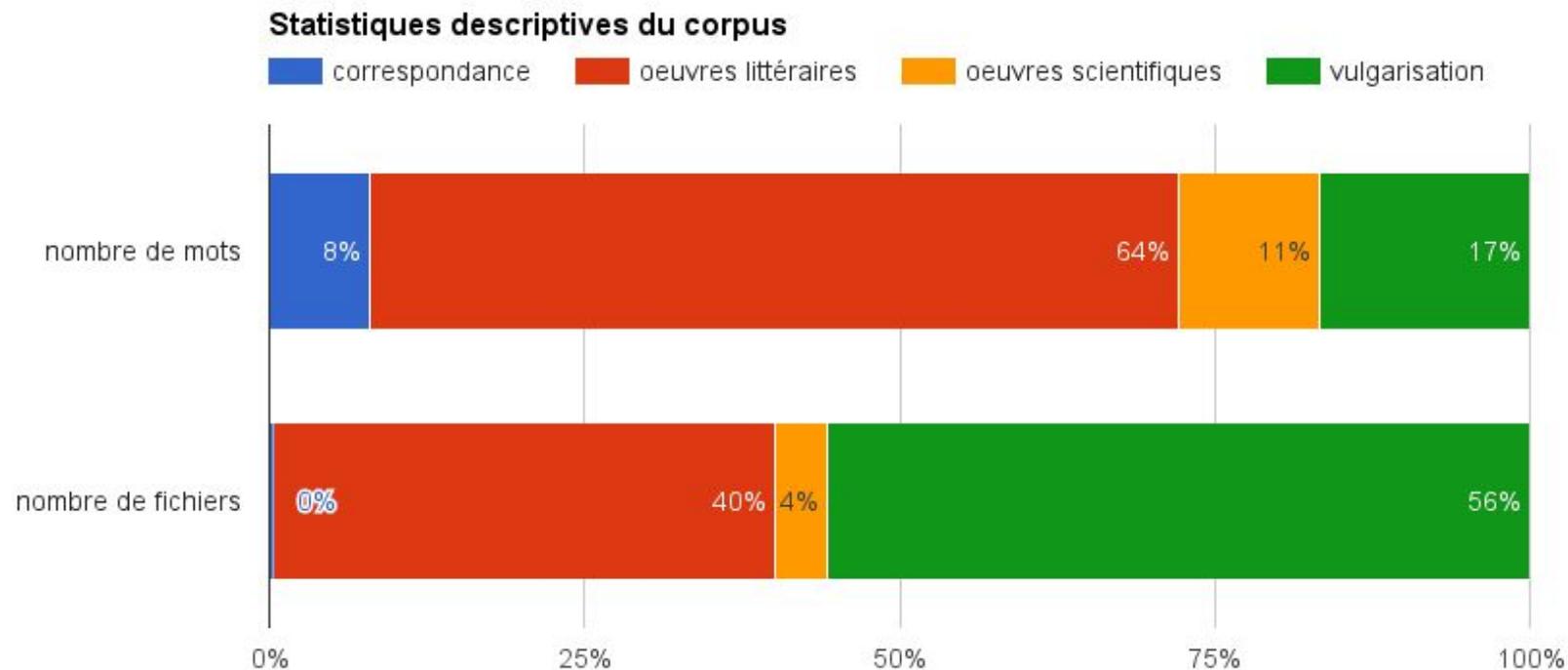
Histoire naturelle, Papillons (1854),
Pellerin

<ark:/12148/btv1b6938165v>



Propriété de l'Éditeur. Déposé

Données quantitatives sur le corpus



Construire le corpus

Henri Coupin (1895) *L'Amateur de papillons, guide pour la chasse, la préparation et la conservation*

<http://gallica.bnf.fr/ark:/12148/bpt6k86300672/f189>



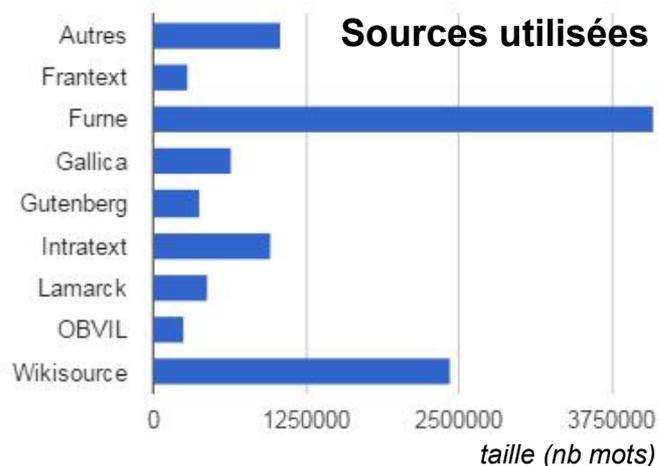
FIG. 163. — Parapluie pour la récolte des papillons.

Construire le corpus : objectifs

- Réunir un grand nombre de textes du XIXe siècle "en lien avec la biologie"
- Faciliter l'exploration visuelle du corpus
- Permettre des études textométriques ciblées sur des sous-corpus
 - *Revue des Deux Mondes*
 - polypes
 - Michelet
- Élargir les possibilités d'analyse :
 - analyse sémantique automatisée (analyse sémantique latente, word2vec, etc.)
 - recherche d'intertextualité
 - étude de la circulation des métaphores biologiques

Construire le corpus : freins et obstacles

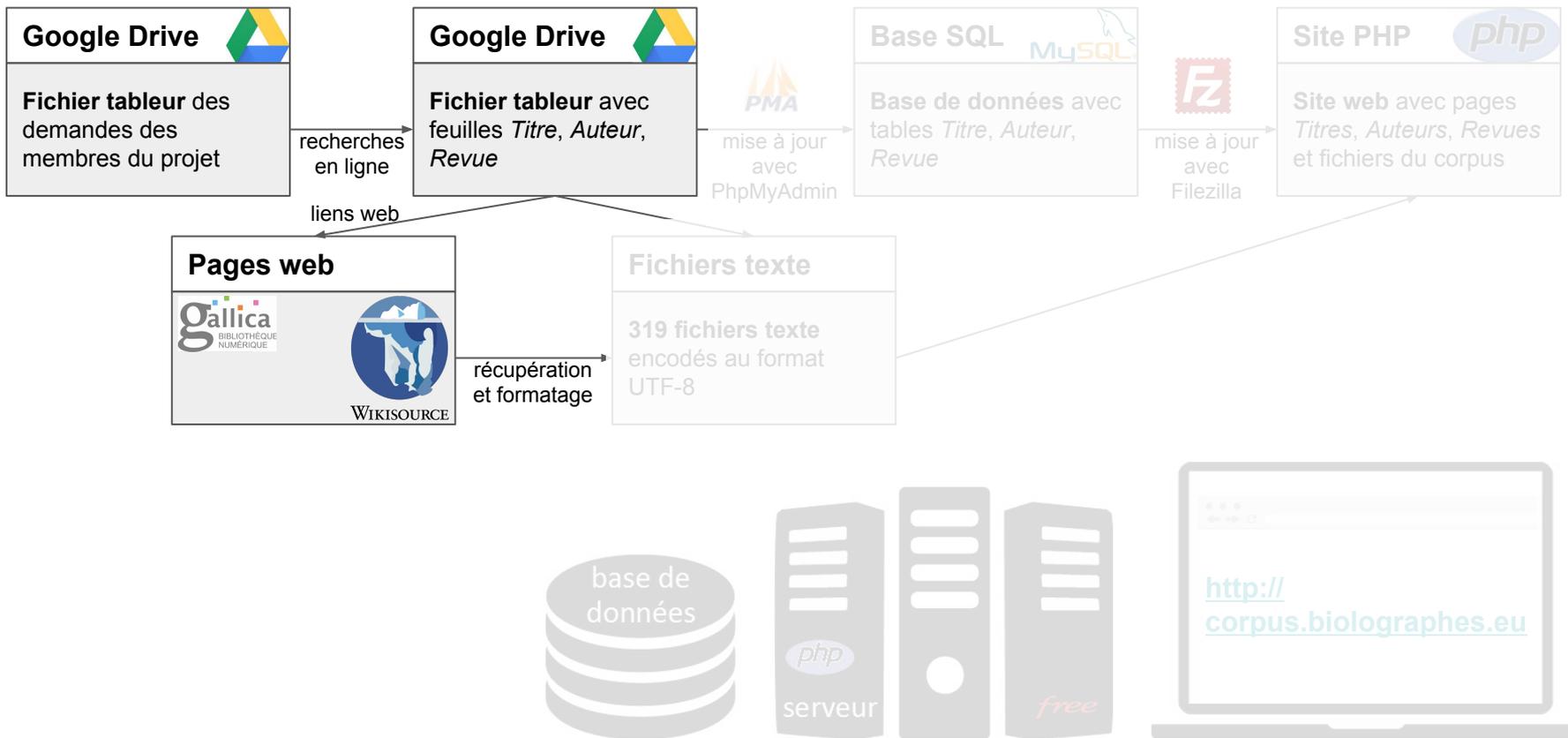
- Choix de l'édition :
idéalement, choix de la dernière édition du vivant de l'auteur
- Textes non disponibles en ligne, ou de mauvaise qualité vs textes disponibles en ligne mais sans informations éditoriales donc non exploitables
(→ métadonnées, TEI)
- Manque d'homogénéité (thématiques, auteurs, etc.) pour le sous-corpus littéraire
- Multilinguisme : choix d'une traduction
(ex. : traduction de Darwin)



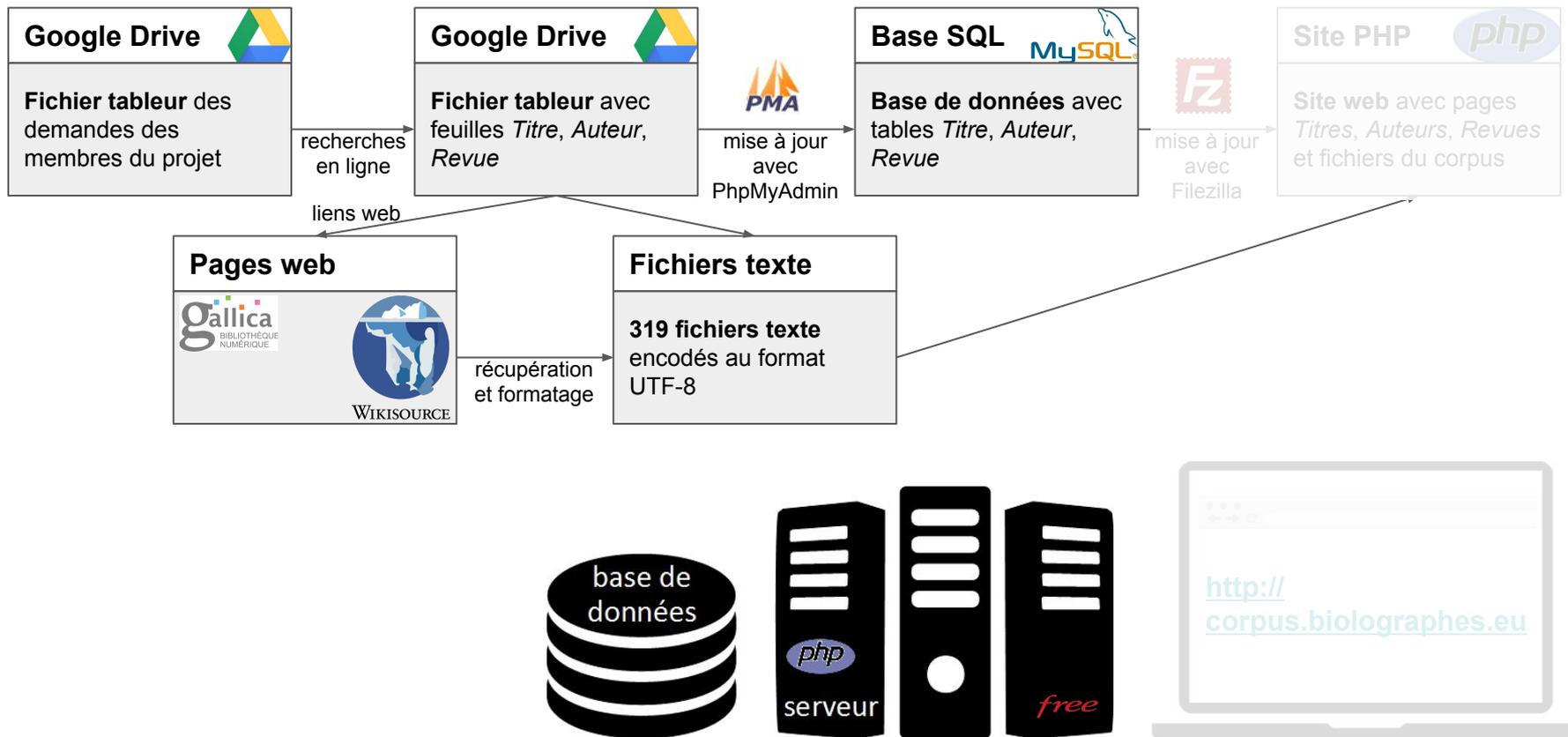
Étapes de construction du corpus



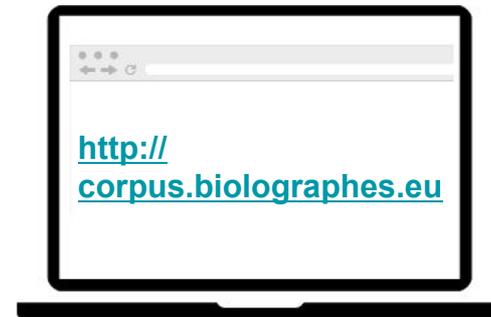
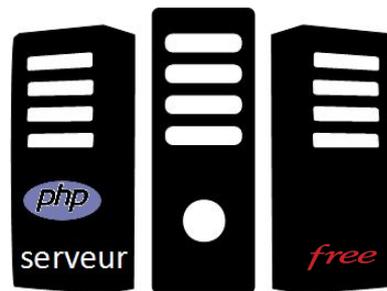
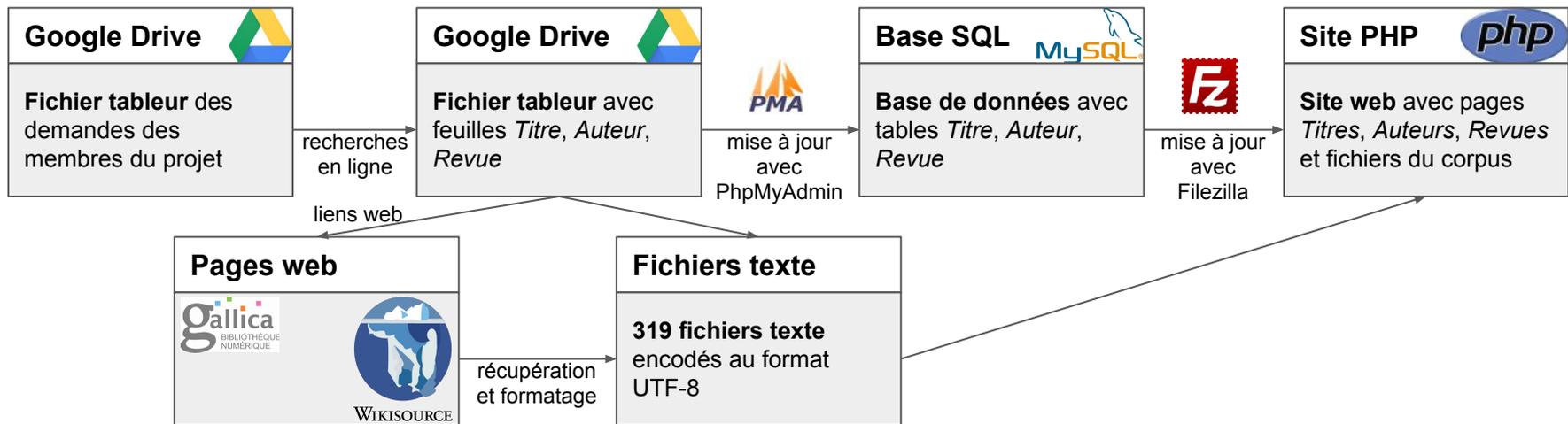
Étapes de construction du corpus



Étapes de construction du corpus

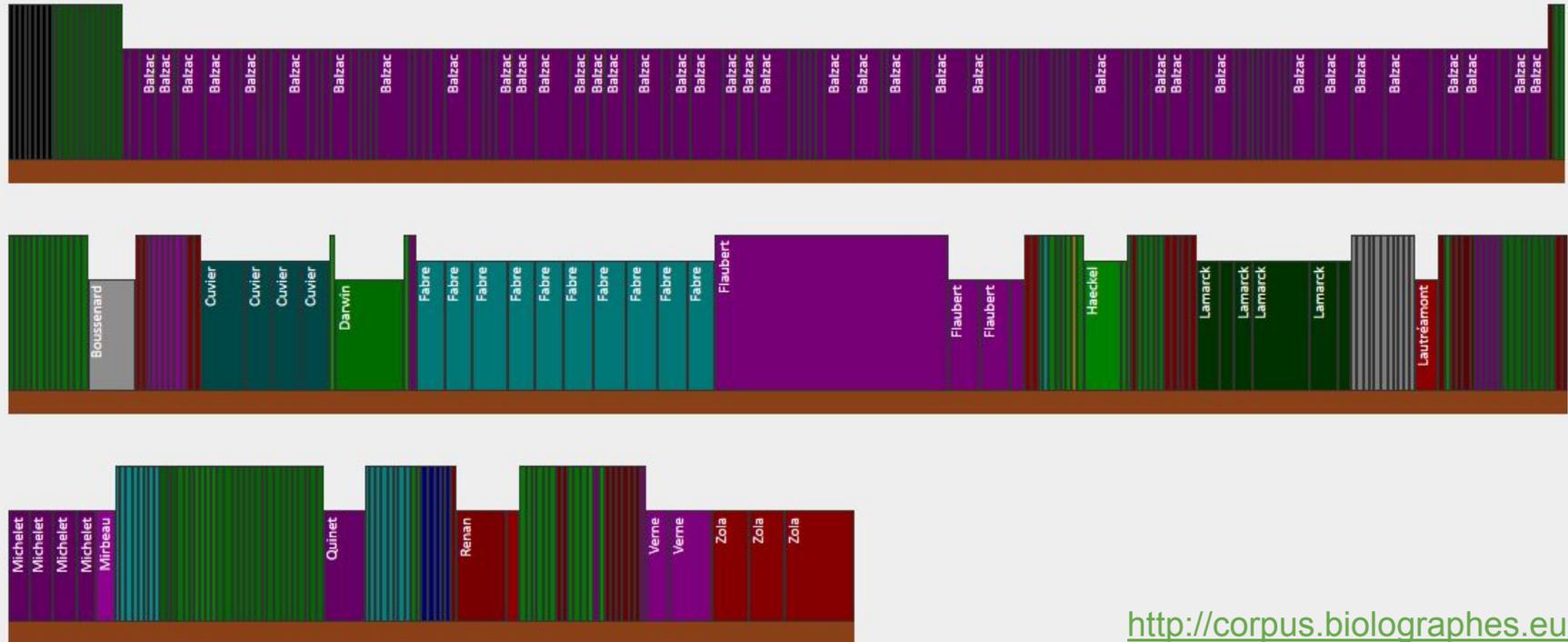


Étapes de construction du corpus



Parcours du site du corpus Biographes

Cliquez sur un ouvrage de la bibliothèque pour accéder à son contenu :



<http://corpus.biographes.eu>

Parcours du site du corpus Biographes

Choix d'un ouvrage



[Accueil](#) [Auteurs](#) [Titres](#) [Revue](#) [À propos](#) [Proposer un ajout au corpus !](#)

Bouvard et Pécuchet

Auteur : [Gustave Flaubert](#)

Référence : Bouvard et Pécuchet - (1881)

Texte intégral sans mise en forme ([source](#)) : [télécharger le fichier OELI-flaubert-bouvard-pecuchet-1910_FR_utf8](#)

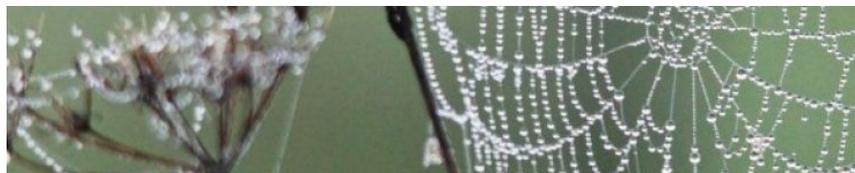
BOUVARD ET PÉCUCHE

ŒUVRE POSTHUME

<http://corpus.biographes.eu>

Parcours du site du corpus Biographes

Page de l'auteur et liste des oeuvres
présentes dans le corpus



[Accueil](#) [Auteurs](#) [Titres](#) [Revue](#) [À propos](#) [Proposer un ajout au corpus](#)

Bouvard et Pécuchet

Auteur : [Gustave Flaubert](#)

Référence : Bouvard et Pécuchet - (1881)

Texte intégral sans mise en forme ([source](#)) : [télécharger le fichier OELI-Flaubert](#)

BOUVARD ET PÉCUCHET

CEUVRE POSTHUME

Gustave Flaubert (1821-1880)

Romancier et auteur dramatique

[Plus d'informations](#)

Œuvres présentes dans le corpus

- [Correspondance 1830-1880](#) - (1830)
- [Bouvard et Pécuchet](#) - (1881)
- [La tentation de Saint Antoine](#) - (1910)
- [Salammbô](#) - (1862)

Parcours du site du corpus Biographes

Plus d'informations : data.bnf.fr

Gustave Flaubert (1821-1880)

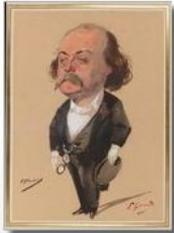
Romancier et auteur dramatique

[Plus d'informations](#)

Œuvres présentes dans le corpus

- [Correspondance 1830-1880](#) - (1830)
- [Bouvard et Pécuchet](#) - (1881)
- [La tentation de Saint Antoine](#) - (1910)
- [Salammbô](#) - (1862)

Gustave Flaubert (1821-1880) (BnF Bibliothèque nationale de France) data.bnf.fr



Pays :	France
Langue :	français
Sexe :	masculin
Naissance :	Rouen, 12-12-1821
Mort :	Croisset (Seine-Maritime), 08-05-1880
Note :	Romancier et auteur dramatique
Domaines :	Littératures
Autre forme du nom :	Giustavas Floberas (1821-1880) (<i>lituanien</i>)
ISNI :	ISNI 0000 0001 2276 2442

Ses activités (624 documents) | Documents à propos de cet auteur | Pages dans data.bnf.fr (2 pages) | Sources et références

Ses activités

Voir tous les documents (624) Voir les documents numérisés (61)

Parcours du site du corpus Biographes

Liste de tous les titres

Corpus primaire

- Balzac, **La Femme abandonnée** (1842)
- Balzac, **Madame Firmiani** (1842)

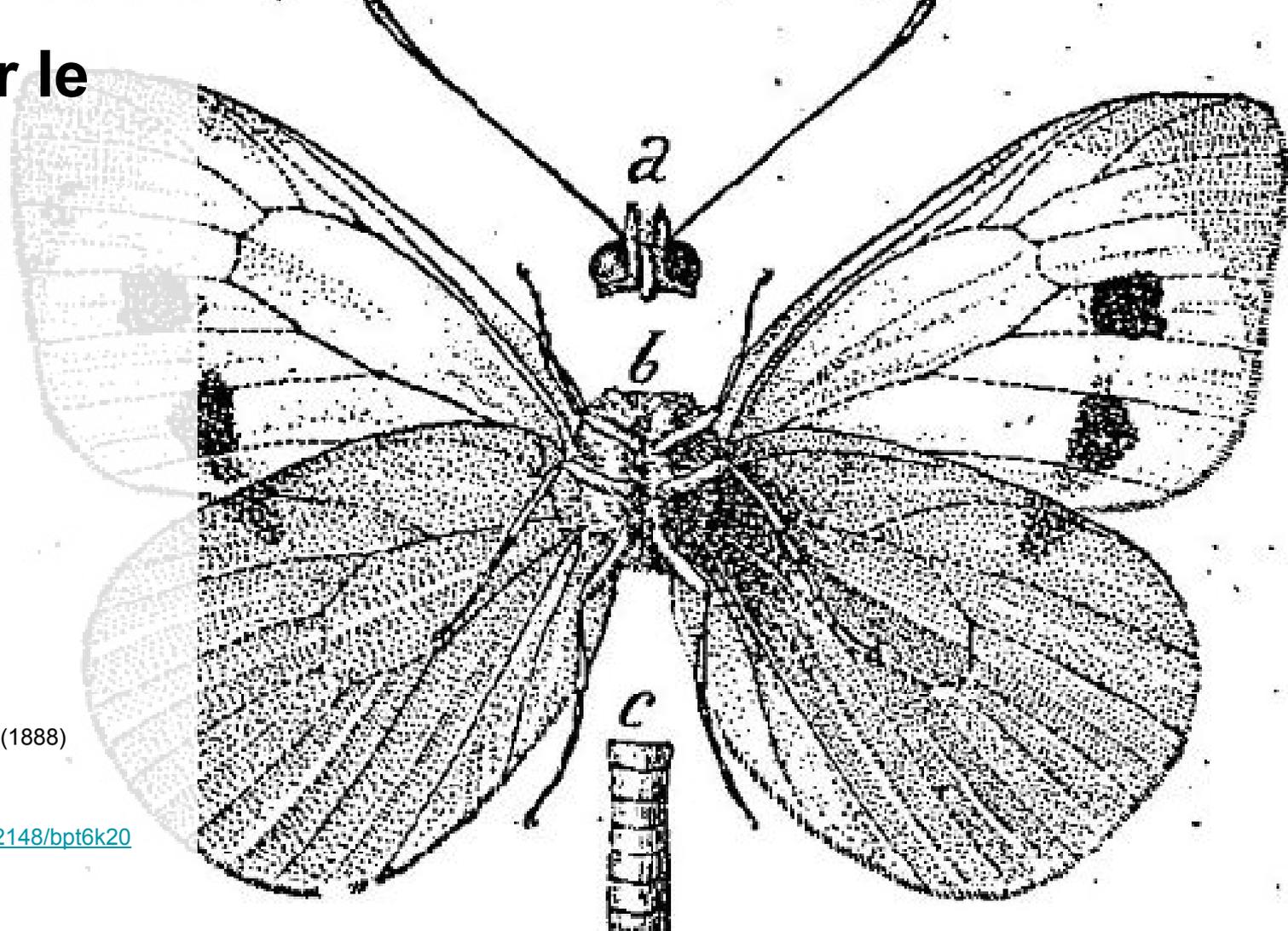
Corpus secondaire 1

- Cuvier, **Tableau élémentaire de l'histoire naturelle des animaux** (1798)
- Cuvier, **Rapport historique sur les progrès des sciences naturelles depuis 1789** (1810)
- Cuvier, **Le Règne animal distribué d'après son organisation, pour servir de base à l'histoire naturelle des animaux et ...** (1800)
- Cuvier, **Discours sur les révolutions de la surface du globe, et sur les changements qu'elles ont produits dans le règne animal** (1800)

Corpus secondaire 2

- Anonyme, **Lettres à un Américain sur l'état des sciences en France III** - *La Revue des deux mondes*, 23, p. 410-437 (1840)
- Anonyme, **Lettres à un Américain sur l'état des sciences en France II** - *La Revue des deux mondes*, 22, p. 532-554 (1840)
- Anonyme, **Lettres à un Américain sur l'état des sciences en France I** - *La Revue des deux mondes*, 21, p. 789-818 (1840)
- Anonyme, **Revue scientifique. Les morts apparentes. L'Académie des Sciences et la planète Leverrier** - *La Revue des deux mondes*, 21, p. 323-332 (1848)
- Anonyme, **Revue scientifique II** - *La Revue des deux mondes*, 21, p. 382-392 (1848)
- Anonyme, **Revue scientifique I** - *La Revue des deux mondes*, 21, p. 324-328 (1848)
- Anonyme, **Les Infiniment petits** - *La Nature. Revue des Sciences*, 4, p. 60-61 (1873)
- Anonyme, **La Vie des animaux par A E Brehm** - *La Nature. Revue des Sciences*, 13, p. 200-203 (1873)
- Babinet, **Télégraphie électrique** - *La Revue des deux mondes*, 2 (1853)

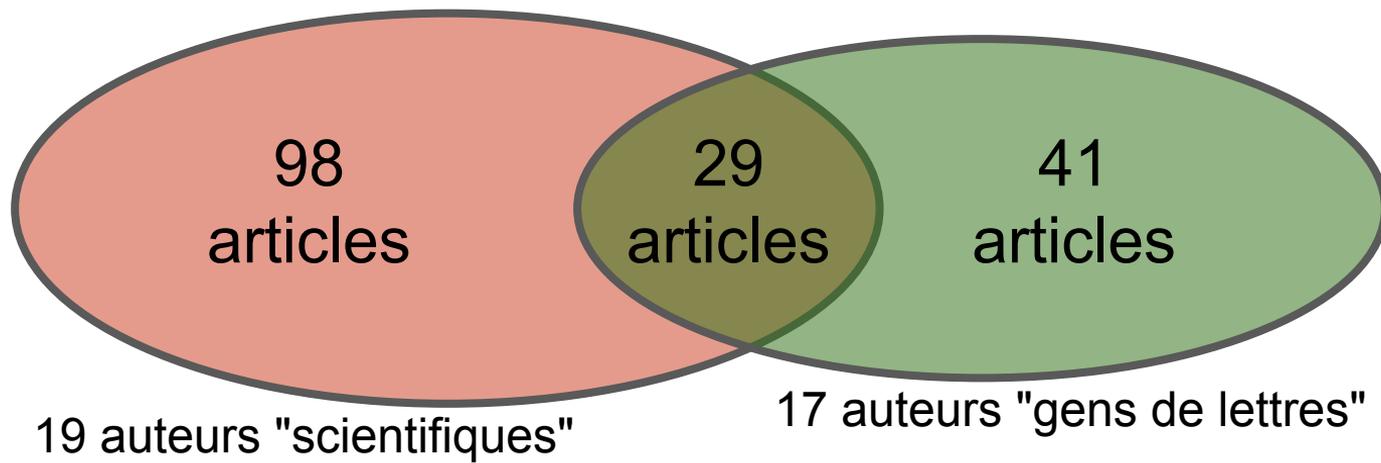
Analyser le corpus



G.-R.-Maurice Maindron (1888)
Les papillons

<http://gallica.bnf.fr/ark:/12148/bpt6k204197j>

Le sous-corpus *Revue des deux Mondes*



Informations quantitatives globales :

- 11 méga-octets
- environ 2 millions d'occurrences, 54 000 formes uniques
- 168 articles

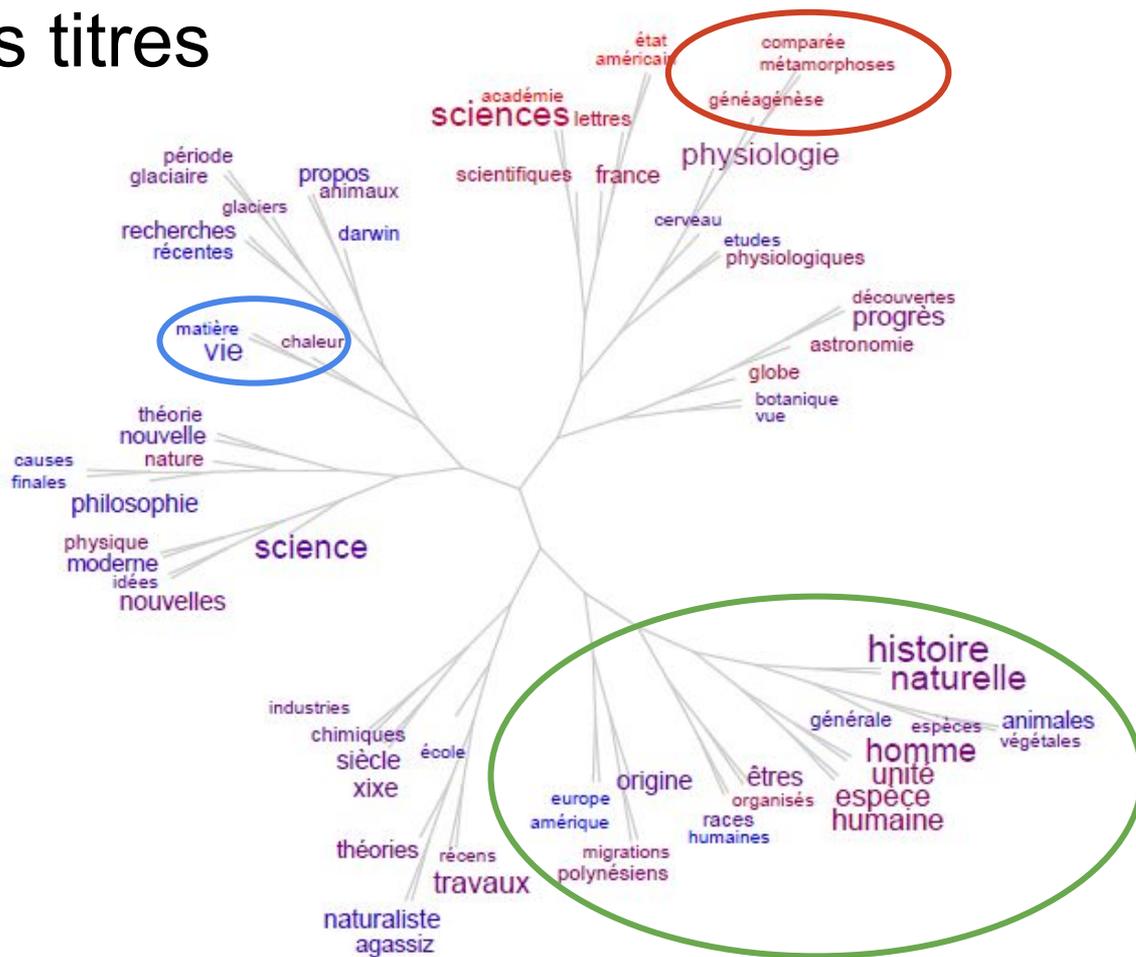
Résumé visuel des titres

Nuage arboré des mots
(hors "mots vides")
présents trois fois ou plus
parmi les titres des articles
du corpus *Biographes*
RDDM

début de siècle

fin de siècle

<http://www.treecloud.org>



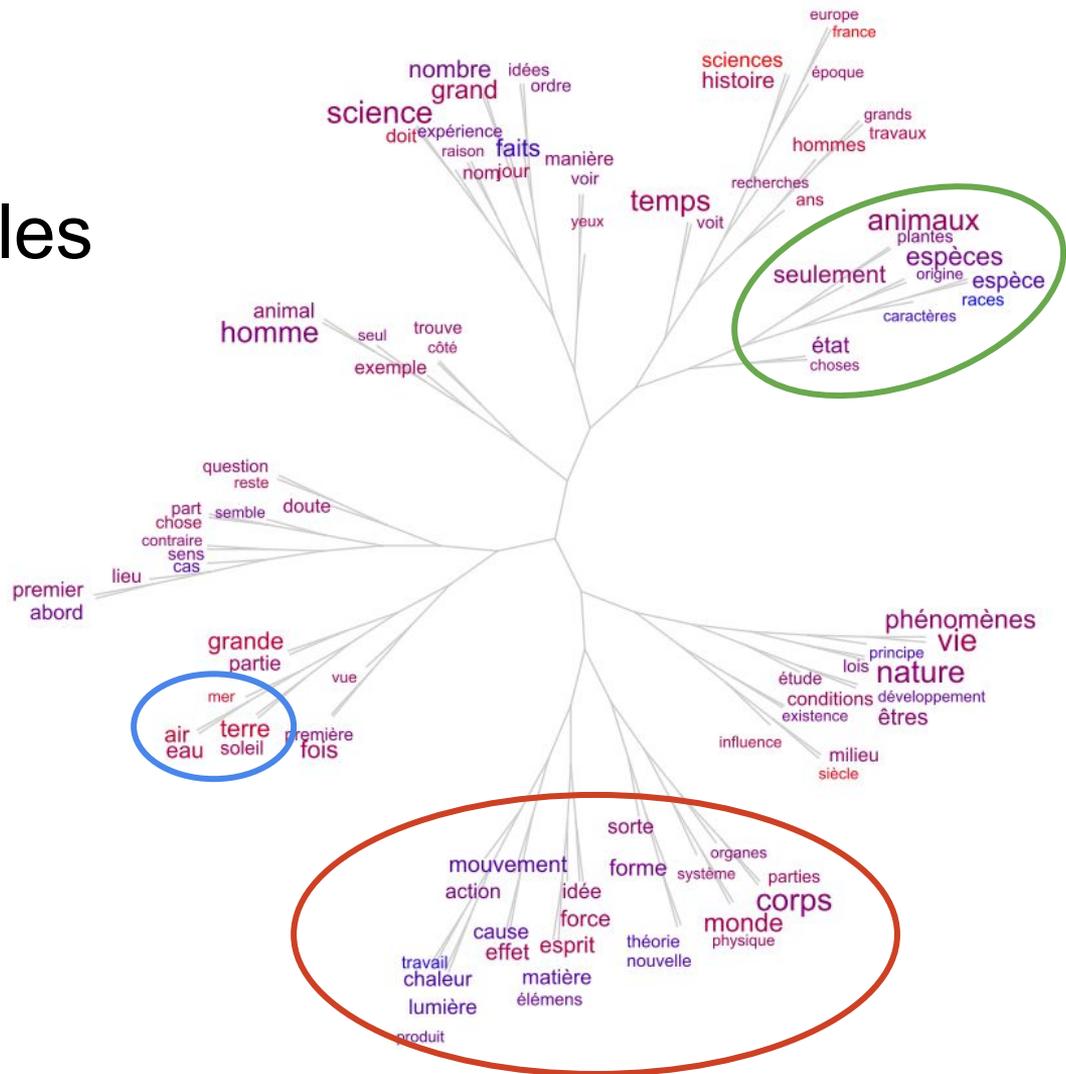
Résumé visuel du contenu des articles

Nuage arboré des 100
mots (hors "mots vides")
les plus fréquents dans les
articles du corpus
Biographes RDDM

début de siècle

fin de siècle

<http://www.treecloud.org>



Vocabulaire des scientifiques et gens de lettres

Chaînes de formes partagées

Laugel & Vacherot

(370) nature
(288) science
(248) vie
(226) homme
(217) philosophie
(207) monde
(197) espèces
(171) phénomènes
(164) principe
(160) êtres
(151) temps
(150) esprit
(148) lois
(144) école
(143) ordre
(140) matière
(134) espèce
(132) animaux
(130) théorie
(127) caractères
(127) loi

Bernard & Martins

phénomènes (324)
corps (268)
vie (244)
animaux (222)
cœur (196)
science (163)
nature (161)
homme (144)
animal (143)
espèces (137)
propriétés (137)
plantes (134)
sciences (131)
sang (128)
organes (122)
végétaux (120)
cerveau (112)
êtres (109)
physiologie (108)
glacier (104)
conditions (100)

Vocabulaire des scientifiques et gens de lettres

Chaînes de formes partagées

Laugel & Vacherot

(370) nature
(288) science
(248) vie
(226) homme
(217) philosophie
(207) monde
(197) espèces
(171) phénomènes
(164) principe
(160) êtres
(151) temps
(150) esprit
(148) lois
(144) école
(143) ordre
(140) matière
(134) espèce
(132) animaux
(130) théorie
(127) caractères
(127) loi

Bernard & Martins

phénomènes (324)
corps (268)
vie (244)
animaux (222)
cœur (196)
science (163)
nature (161)
homme (144)
animal (143)
espèces (137)
propriétés (137)
plantes (134)
sciences (131)
sang (128)
organes (122)
végétaux (120)
cerveau (112)
êtres (109)
physiologie (108)
glacier (104)
conditions (100)

Vocabulaire des scientifiques et gens de lettres

Gens de lettres

(905) vie

(869) nature

(822) science

(672) homme

(633) corps

(539) conscience

(518) esprit

(475) monde

(466) force

(465) phénomènes

Scientifiques

animaux (1190)

espèces (1161)

corps (1143)

temps (1082)

homme (1062)

vie (1052)

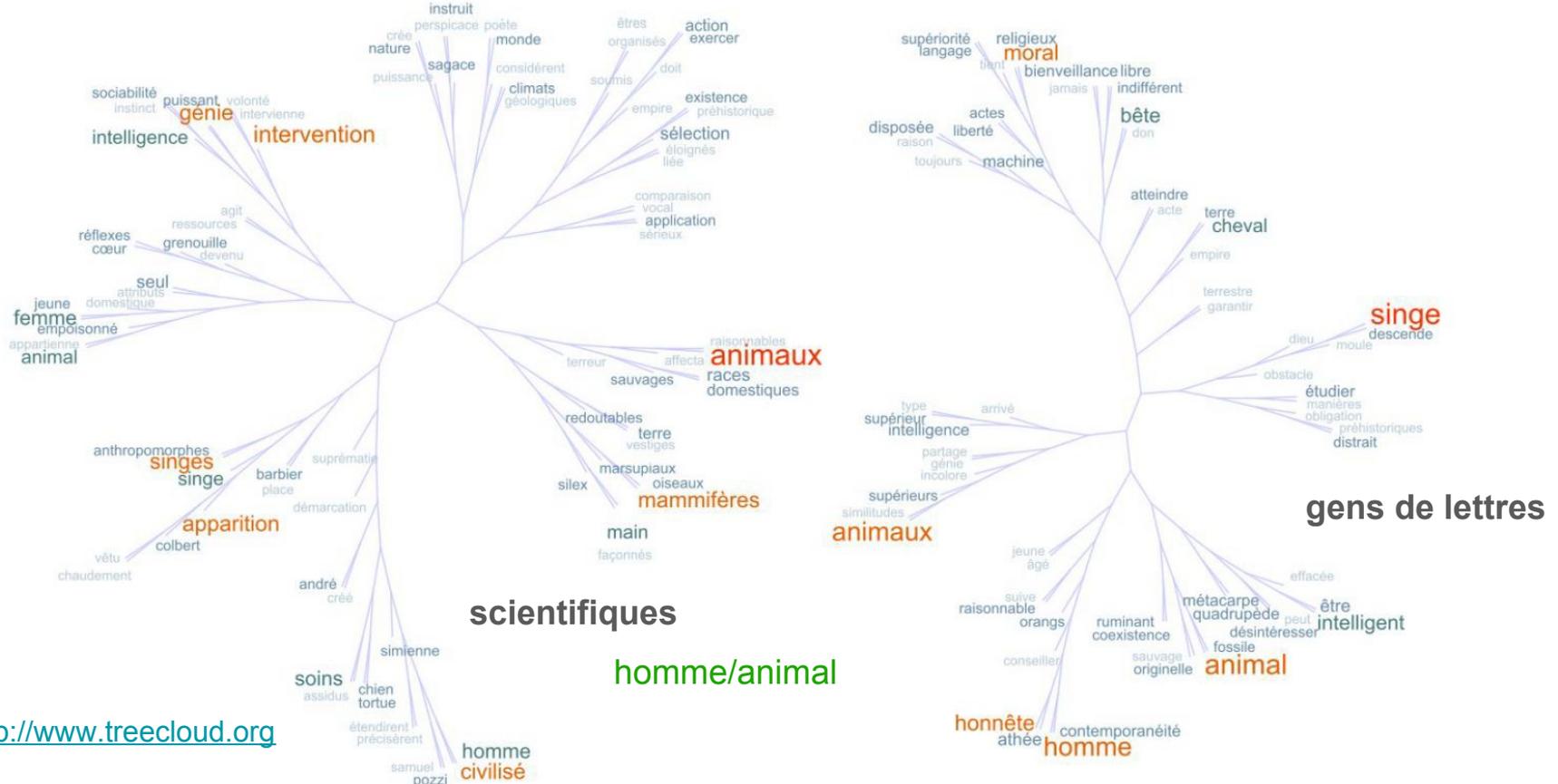
nature (1012)

air (962)

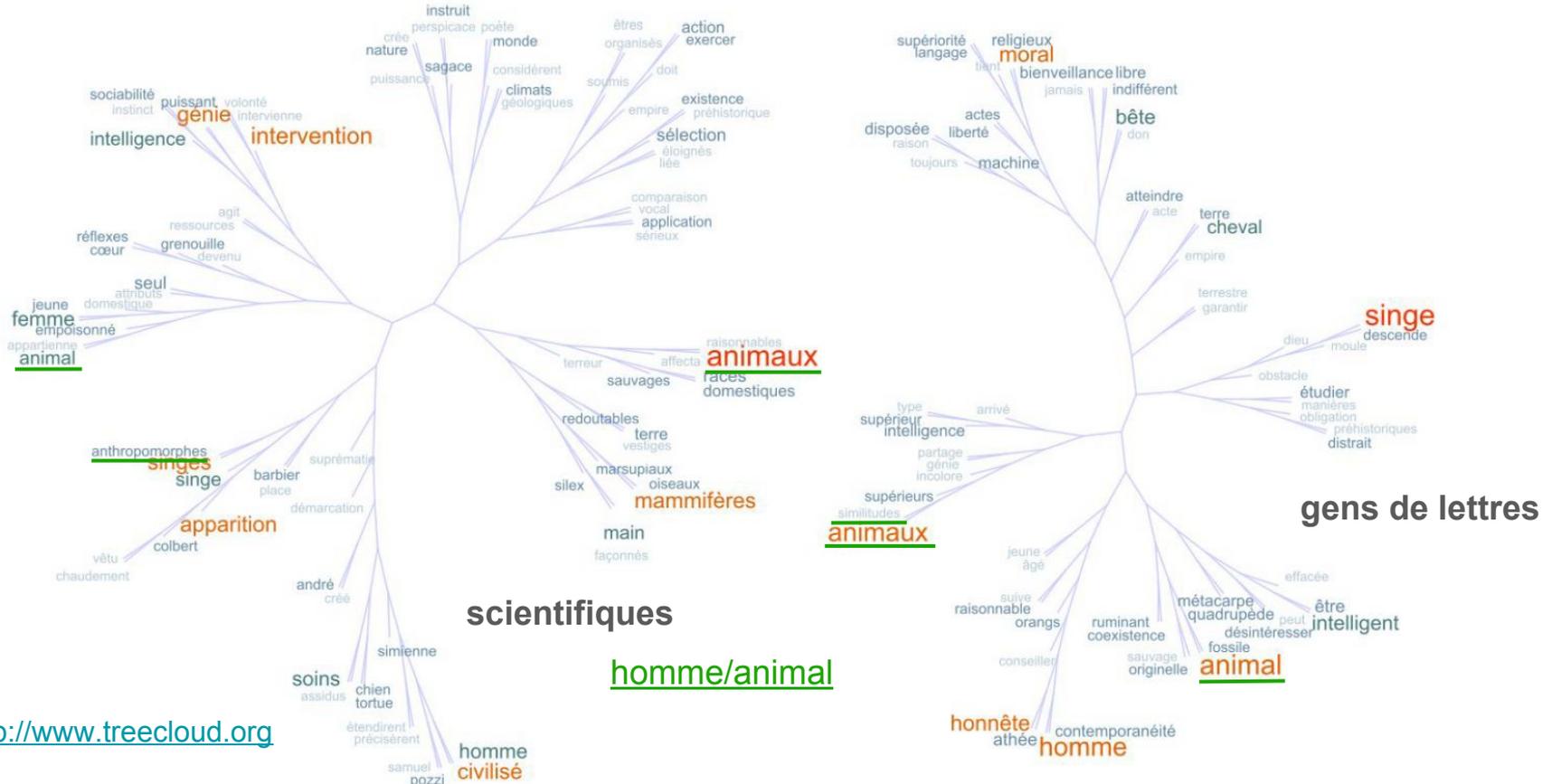
science (954)

eau (950)

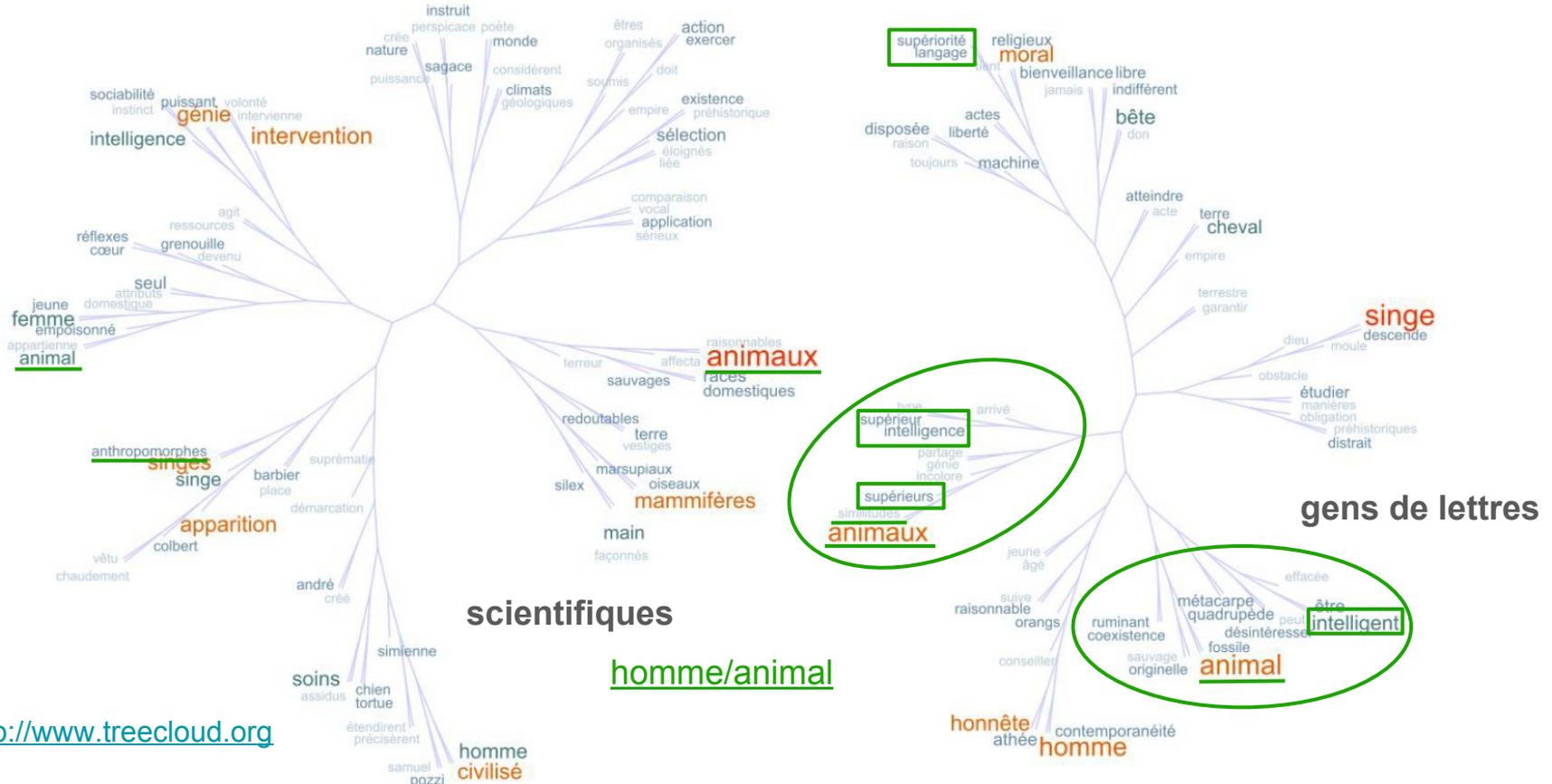
Nuages arborés des contextes de « homme »



Nuages arborés des contextes de « homme »

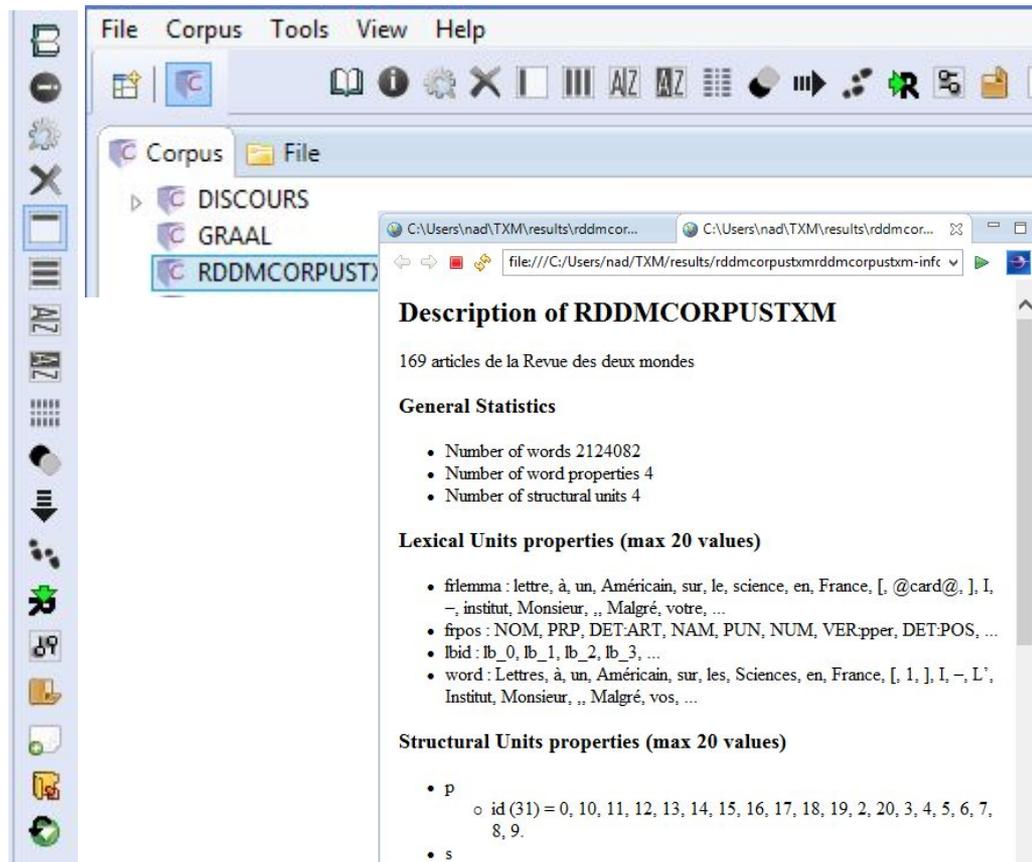


Nuages arborés des contextes de « homme »



TXM

- Étiquetage des données
- Lemmatisation
- Édition des textes
- Partition du corpus
- Spécificités
- **Contextes**
- Étude du lexique
- Recherche de motifs linguistiques (CQL)
- Recherches contextuelles



The screenshot displays the TXM software interface. The main window shows a file explorer view of a corpus named 'RDDMCORPUSTXM'. A secondary window titled 'Description of RDDMCORPUSTXM' is open, providing detailed statistics and properties for the corpus.

Description of RDDMCORPUSTXM

169 articles de la Revue des deux mondes

General Statistics

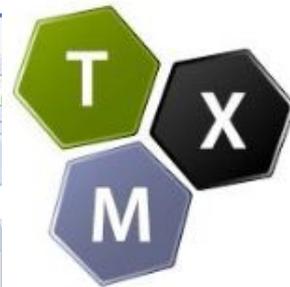
- Number of words 2124082
- Number of word properties 4
- Number of structural units 4

Lexical Units properties (max 20 values)

- frlemma : lettre, à, un, Américain, sur, le, science, en, France, [, @card@,], I, -, institut, Monsieur, ,, Malgré, votre, ...
- frpos : NOM, PRP, DET:ART, NAM, PUN, NUM, VER:pper, DET:POS, ...
- lbid : lb_0, lb_1, lb_2, lb_3, ...
- word : Lettres, à, un, Américain, sur, les, Sciences, en, France, [, 1,], I, -, L', Institut, Monsieur, ,, Malgré, vos, ...

Structural Units properties (max 20 values)

- p
 - id (31) = 0, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 20, 3, 4, 5, 6, 7, 8, 9.
- s
 - n (782) = 1, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 20, 3, 4, 5, 6, 7, 8, 9.



Accès au texte



qui a vécu du fonds d'idées qu'il s'était fait dans son Discours sur l'origine de l'inégalité, Buffon a constamment travaillé sur lui-même et recommencé, d'année en année, son éducation scientifique. Aussi son Histoire naturelle manque-t-elle un peu d'ordre et d'unité ; les parties n'en semblent pas avoir de justes proportions entre elles ; et sous l'assurance qu'il affecte ou dont la fermeté de son style est l'expression extérieure, la vérité est que Buffon ne sait bien souvent quel choix faire entre la diversité des hypothèses que la fécondité de son imagination lui suggère. Pour ne parler ici que d'une seule question, Flourens n'hésitait pas à en faire le partisan décidé de la fuite des espèces, et à l'appui de son dire il abondait en citations topiques. Mais M. de Lanessan n'en a pas apporté de moins nombreuses ni de moins caractéristiques pour prouver qu'au contraire, avant Lamarck et Darwin, Buffon avait conçu la doctrine de l'indéfinie variabilité des espèces. Et cela, si je ne me trompe, signifie deux choses à la fois : la première, que Buffon, sur cette question comme sur bien d'autres, a longtemps ou toujours hésité ; et la seconde, que, puisque l'on peut le réclamer pour soi des deux parts, c'est que l'étendue de son regard avait d'abord embrassé l'horizon de la question tout entière. Nous n'avons point ici qualité pour le juger comme

Index et fréquence des termes



RDDMCORPUSTXM: [frpos="NAM"];word

Query: [frpos="NAM"]

Thresholds: Fmin: 1 Fmax: 9999999 Vmax: 9999999 Page size: 100

1 - 100 / 7942

t 40562 , v 7942 , fmin 1 , fmax 759

word	Frequency	T=2124082
Galilée	235	
Lamarck	227	
Cuvier	201	
Allemagne	186	
M...	176	

RDDMCORPUSTXM: [word="Lamarck"]

Query: [word="Lamarck"]

sort keys: #1 None #2 None #3 None #4 None Sort

1 - 100 / 228

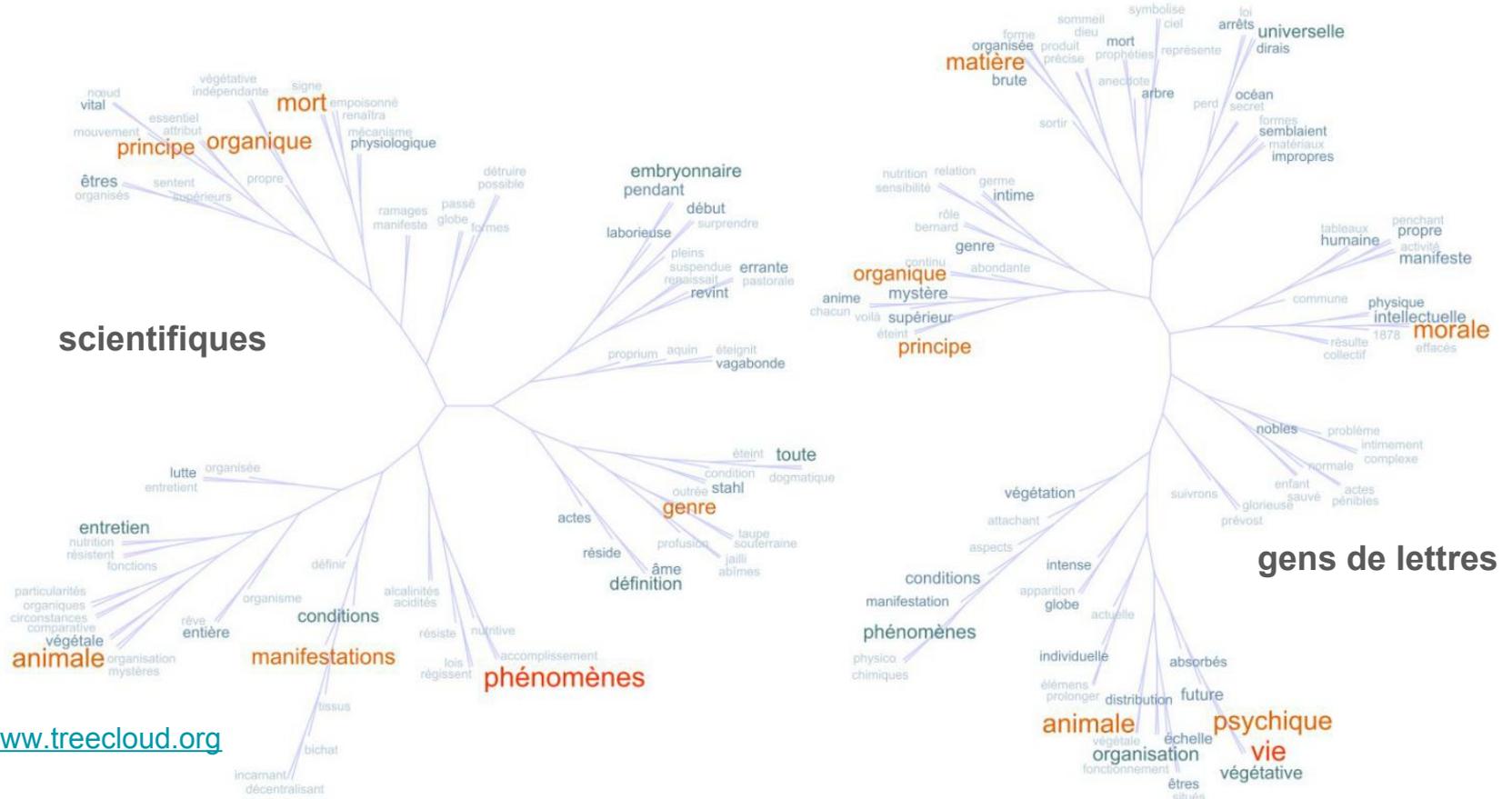
Hide settings

text_id	Left context	Keyword	Right context
ARRDM_brunetiere_1888-89	caractéristique pour prouver qu'au contraire, avant	Lamarck	et Darwin, Buffon avait conçu la doctrine de l'indéfinie va
ARRDM_brunetiere_1888-89	les naturalistes qui l'ont suivi, les	Lamarck	, les Cuvier, les Geoffroy Saint-Hilaire, pour ne nommer qu
ARRDM_brunetiere_1892-113bis	dirait-on pas déjà du Buffon, ou du	Lamarck	, ou du Darwin ? On pourrait s'y tromper. En
ARRDM_fouillee_1979-34-II	deviendra instinct, selon la profonde doctrine de	Lamarck	, de Darwin et de Spencer. Le corps social, lui
ARRDM_hilaire_1837-10	être associés l'un à l'autre,	Lamarck	et Cuvier. La longue et honorable vie de Lamarck se divise

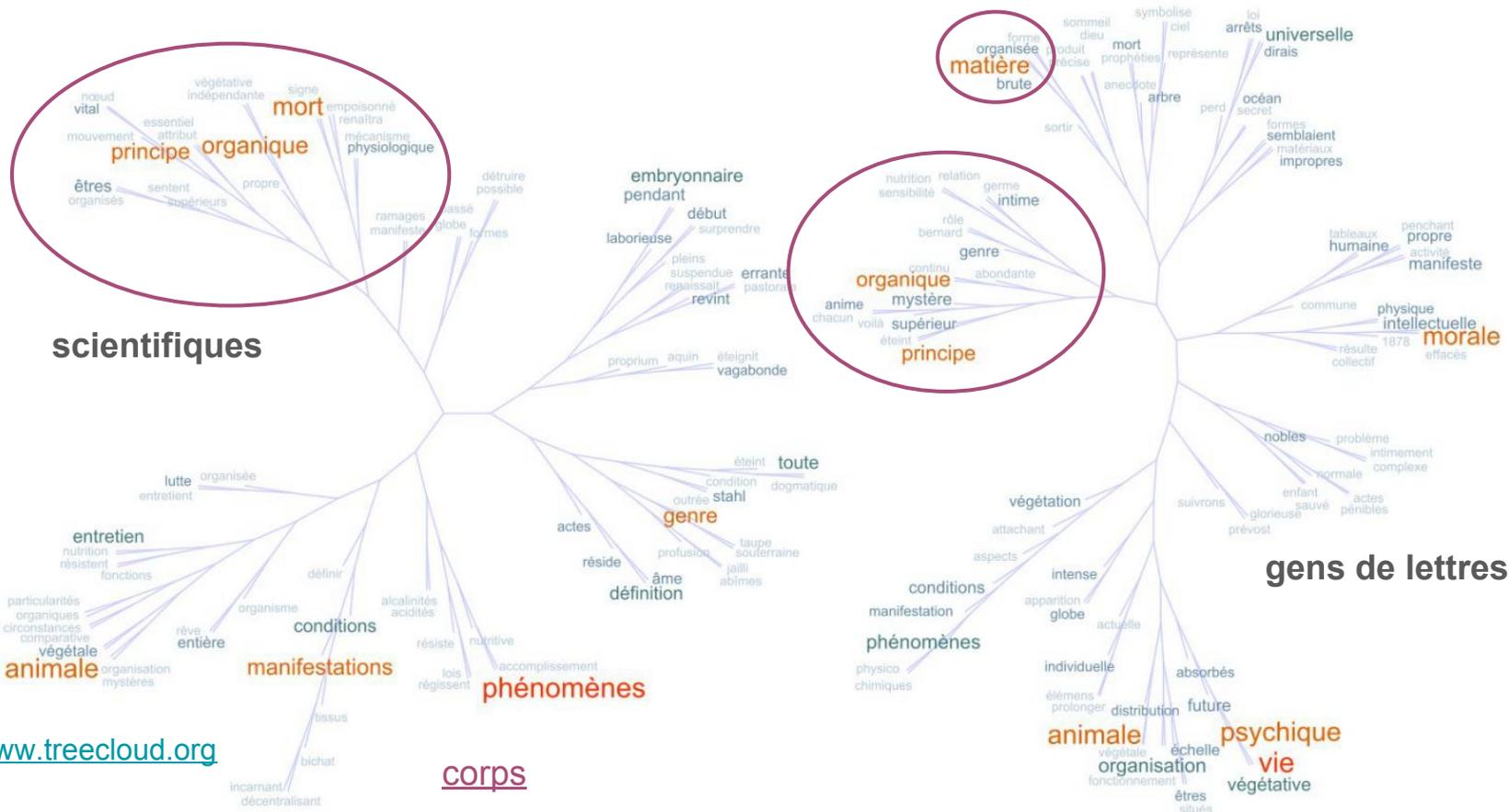
Concordancier



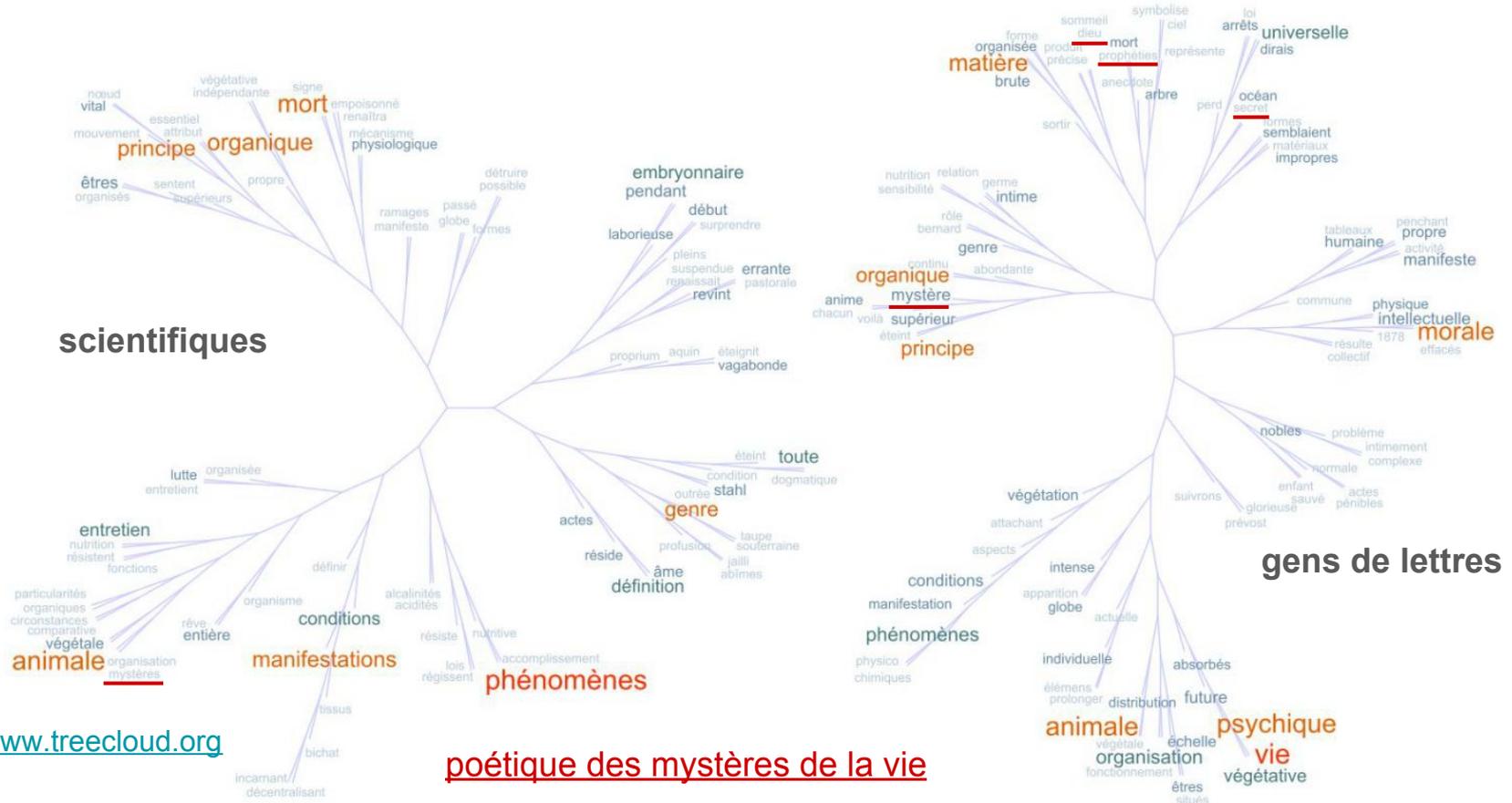
Nuages arborés des contextes de « vie »



Nuages arborés des contextes de « vie »



Nuages arborés des contextes de « vie »



<http://www.treecloud.org>

poétique des mystères de la vie

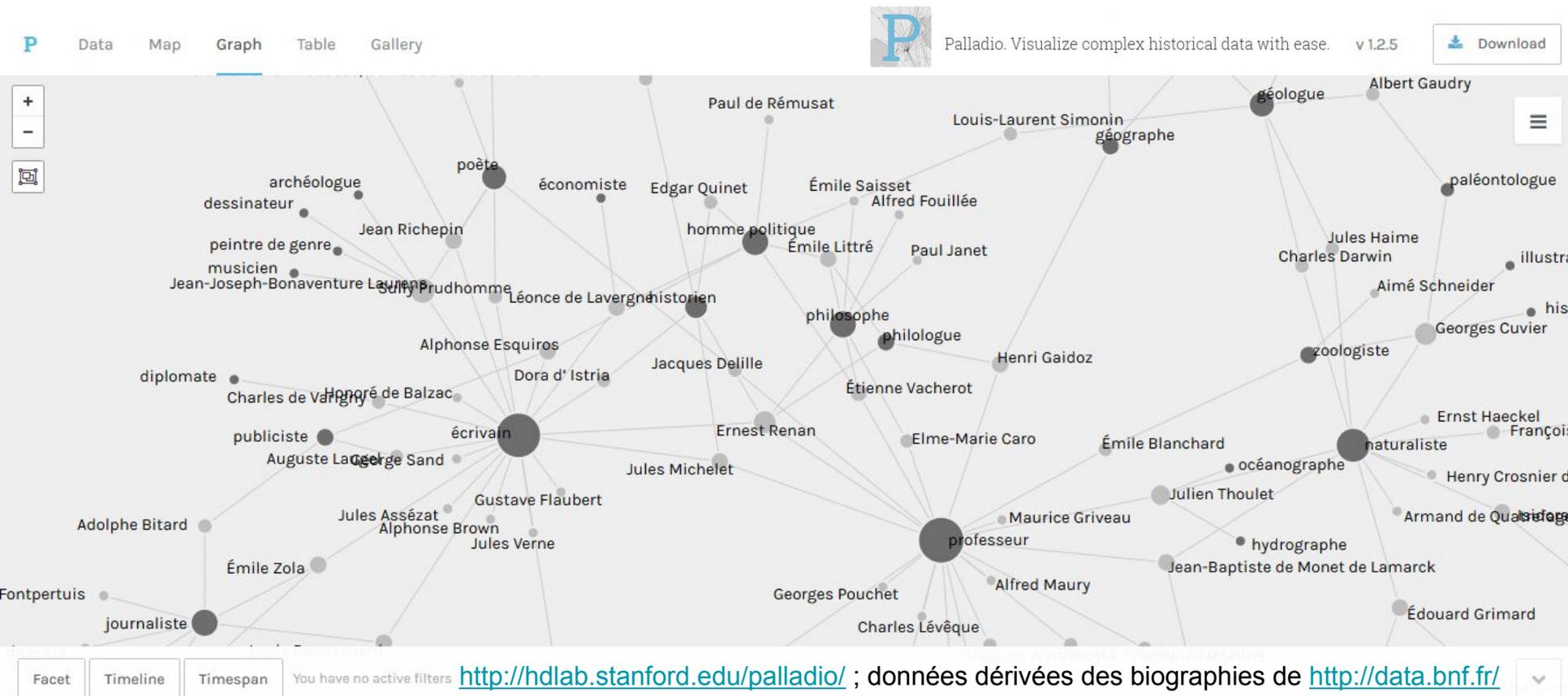
Cartographier le corpus



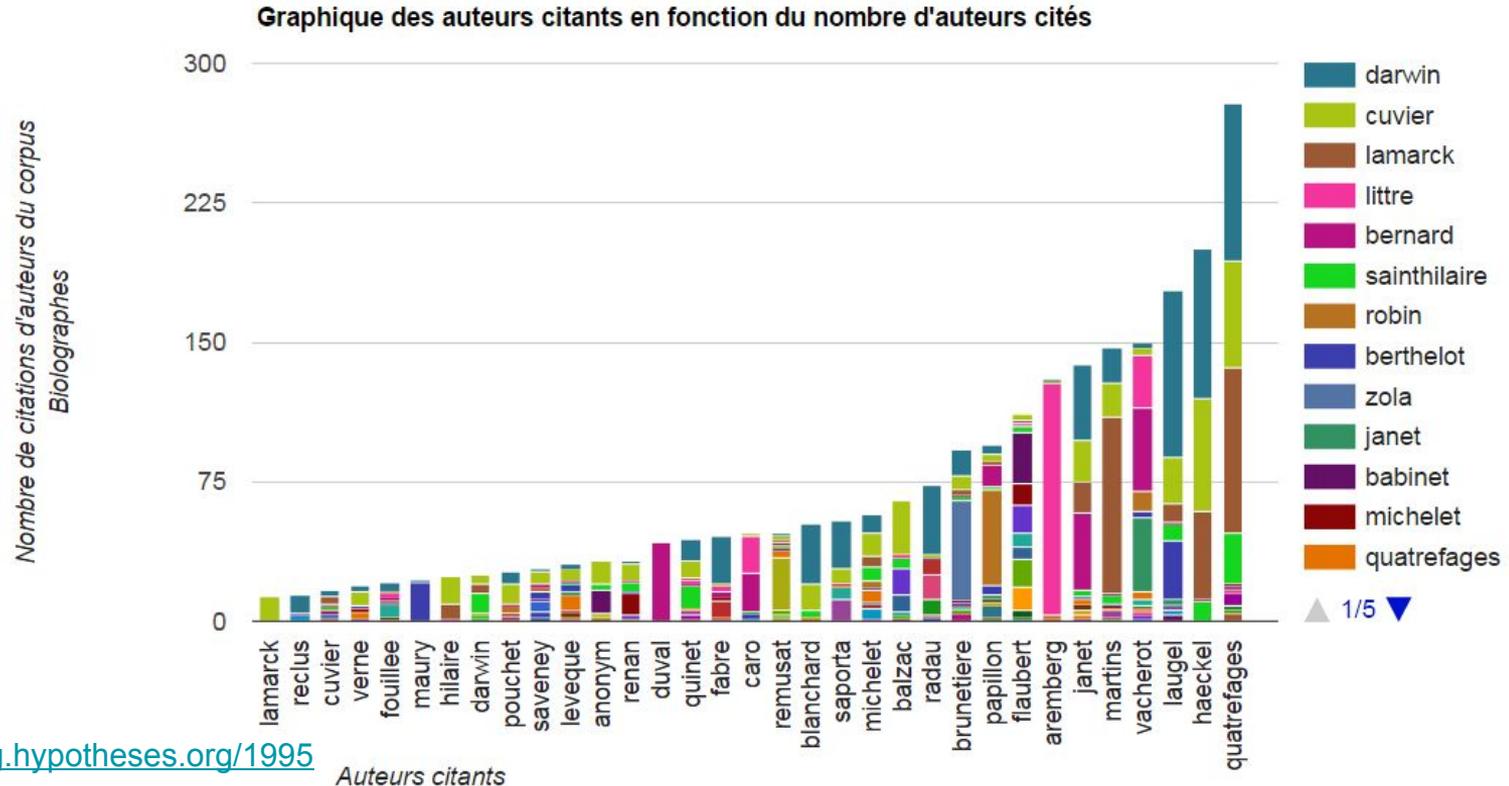
Amédée Varin (1862) *Les papillons, métamorphoses terrestres des peuples de l'air*

<http://gallica.bnf.fr/ark:/12148/bpt6k1228258/f1>

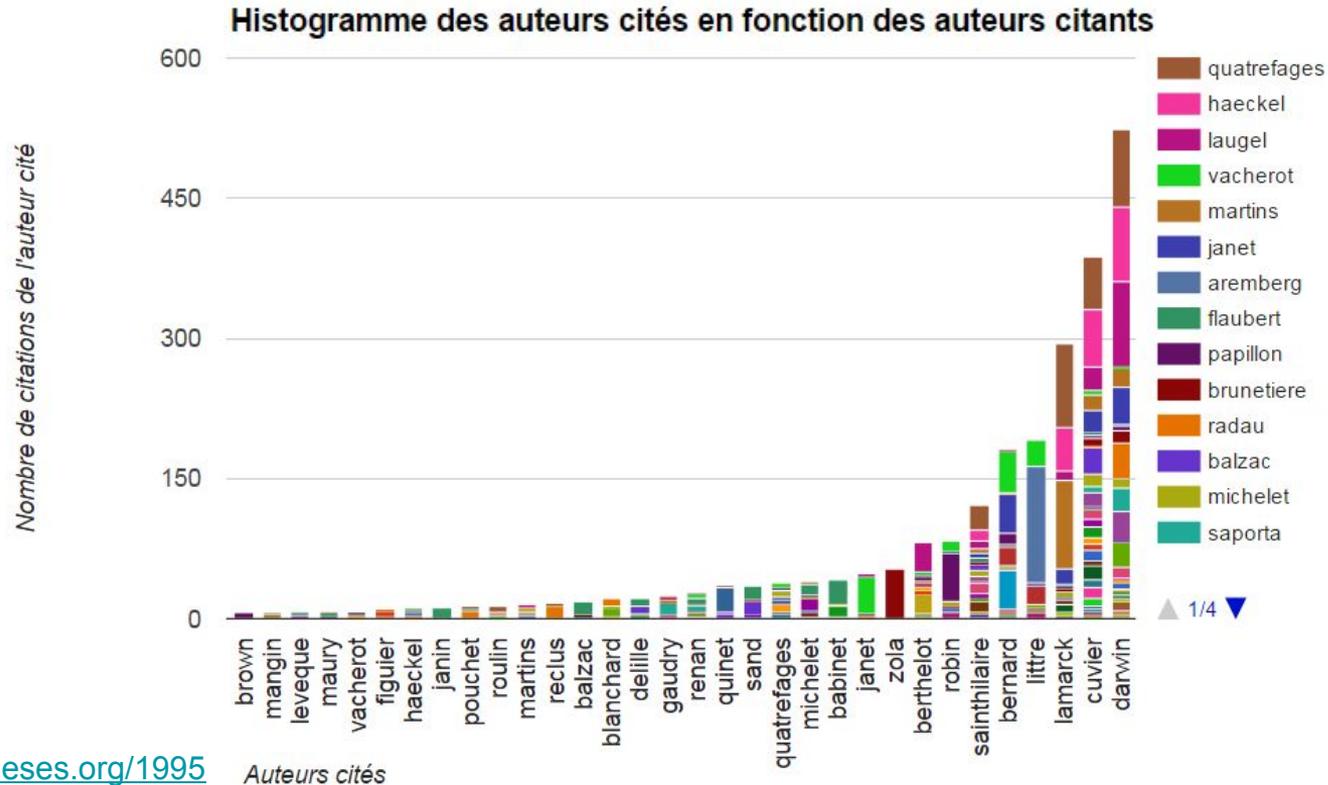
Données sur les auteurs



Visualisation des citations entre auteurs du corpus



Visualisation des citations entre auteurs du corpus



Visualisation des citations

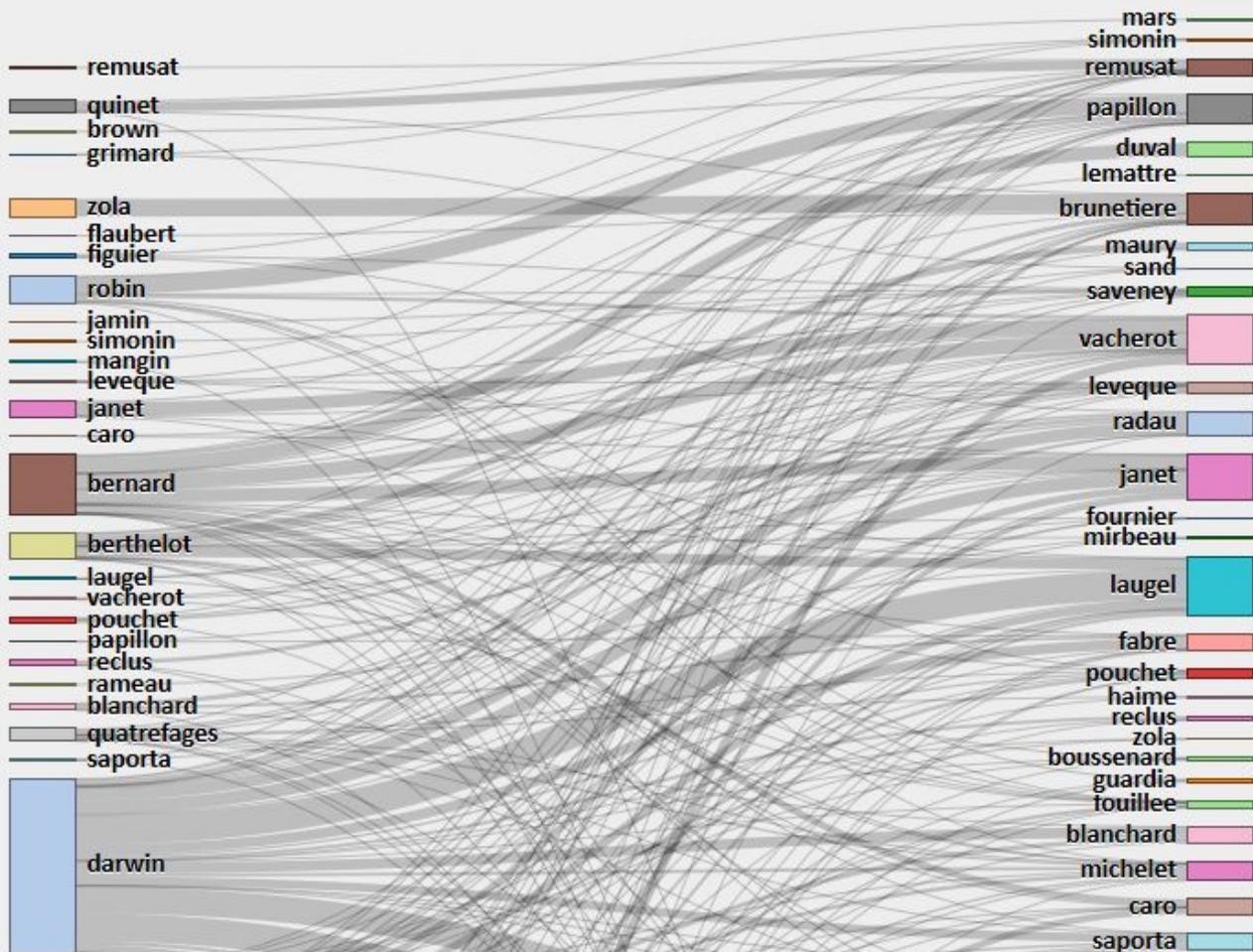
<https://biolog.hypotheses.org/1995>

auteurs cités

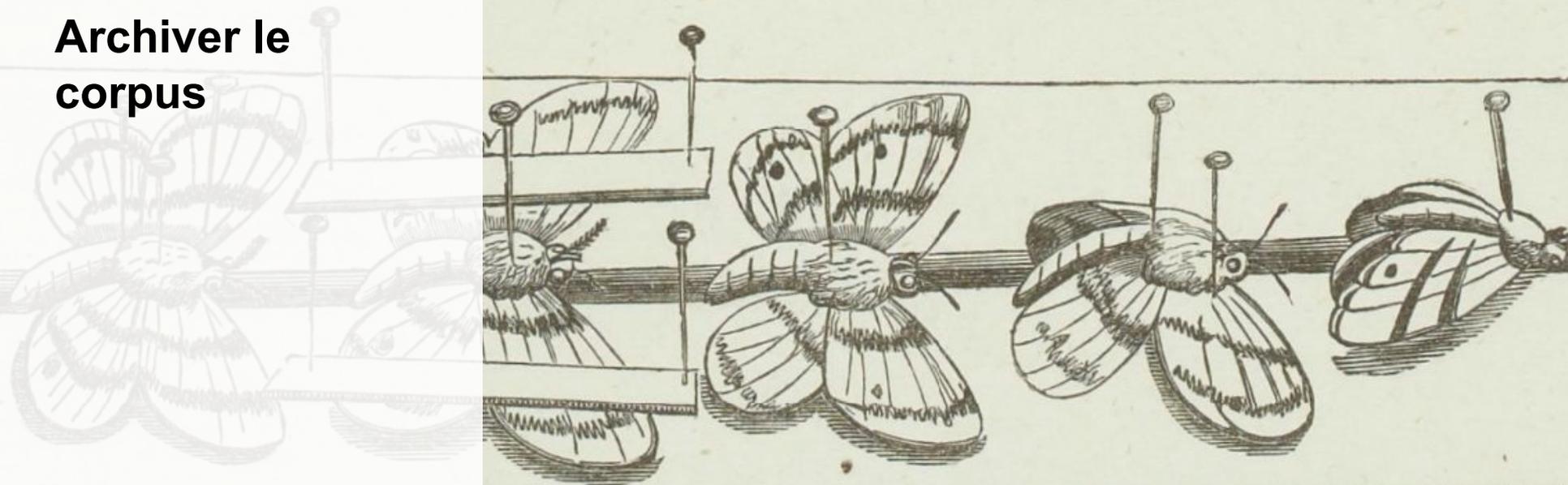
remusat
quinet
brown
grimard
zola
flaubert
figuier
robin
jamin
simonin
mangin
leveque
janet
caro
bernard
berthelot
laugel
vacherot
pouchet
papillon
reclus
rameau
blanchard
quatrefoies
saporta
darwin

auteurs citants

mars
simonin
remusat
papillon
duval
lemattre
brunetiere
maury
sand
saveney
vacherot
leveque
radau
janet
fournier
mirbeau
laugel
fabre
pouchet
haimé
reclus
zola
bousenard
guardia
touillee
blanchard
michelet
caro
saporta



Archiver le corpus



Amédée Varin (1862) *Les papillons,
métamorphoses terrestres des
peuples de l'air*

<http://gallica.bnf.fr/ark:/12148/btv1b86002237/f279>

Archivage avec HumaNum



Huma-Num : la très grande infrastructure des humanités numériques

Mise à disposition pérenne du corpus Biographes sous 2 formes :

- Liste de fichiers XML-TEI **téléchargeables** (avec métadonnées)
- Fichiers **interrogeables** avec l'outil Philologic :
collaboration en cours avec les équipes d'HumaNum
pour l'installation de pré-versions de Philologic 4



Pre-release

v4.5-rc6
c0030b1

Philologic 4.5rc6

clovis released this 2 days ago · 9 commits to master since this release

Bug fixes

Downloads

Source code (zip)

Source code (tar.gz)

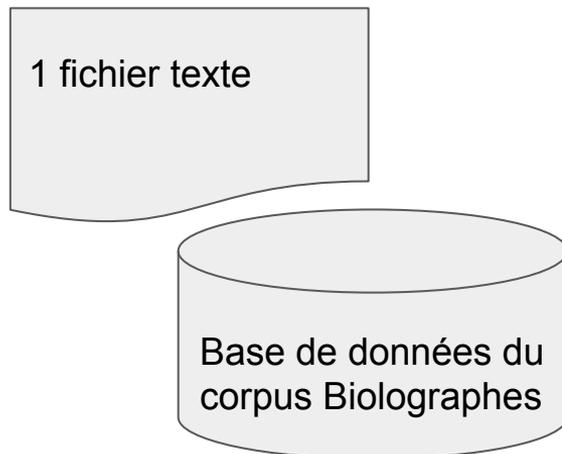
<https://github.com/ARTFL-Project/PhiloLogic4>

Démo sur <http://artflsrv02.uchicago.edu/philologic4/encyclopedie/>

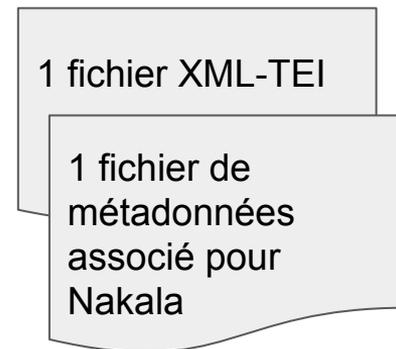
Formatage des données pour Nakala



Entrée :



Sortie pour Nakala :



Format Nakala

1 fichier XML-TEI

```
<TEI version="5.0"
xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt>
<title>Les Infiniment petits</title>
<author>Anonyme</author>
<date>1873</date>
</titleStmt>
<sourceDesc>https://fr.wikisource.org/
wiki/Les\_Infiniment\_petits
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>LES INFINIMENT
PETITS
```

1 fichier de métadonnées associé

```
<nkl:Data xmlns:nkl="http://nakala.fr/schema#"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dcterms="http://purl.org/dc/terms/"
xsi:schemaLocation="http://purl.org/dc/terms/
http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd">
<dcterms:title>Les Infiniment petits</dcterms:title>
<dcterms:creator>Lechevrel, Nadège</dcterms:creator>
<dcterms:type>Text</dcterms:type>
<dcterms:created>2016</dcterms:created>
<dcterms:identifiant>https://fr.wikisource.org/wiki/Les\_Infiniment\_petits
</dcterms:identifiant>
<dcterms:subject>[SHS:LIT] Humanities and Social Sciences/Literature
</dcterms:subject>
<dcterms:rights>Public Domain</dcterms:rights>
<dcterms:licence xsi:type="dcterms:URI">
https://creativecommons.org/publicdomain/mark/1.0/
</dcterms:licence>
<dcterms:source>ANR-13-FRAL-0013</dcterms:source>
<nkl:inCollection>11280/685b4329</nkl:inCollection>
<nkl:dataFormat>TXT UTF-8</nkl:dataFormat>
</nkl:Data>
```

Import dans Nakalona (Nakala + Omeka) nakalona

Requête : biologaphes Type de requête : Mot-Clé

Types d'enregistrement : Contenu, Fichier, Collection, Exposition, Page de l'exposition, Simple Page

Types d'enregistrement	Titre
Collection	nakala.biographes

Propulsé par Omeka | Documentation | Forums de support Version 2.3 | Informations système

Dublin Core

Titre nakala.biographes

Identifiant 11280/685b4329

Contenus ajoutés récemment

- 14 sept. 2016 Les Infiniment petits
- 14 sept. 2016 Histoire de la création des êtres organisés d'après les lois naturelles...
- 14 sept. 2016 Anthropogénie, ou Histoire de l'évolution humaine : leçons familières sur les principes de l'embryologie et de la phylogénie humaines

[Modifier](#)

[Voir la page publique](#)

[Supprimer](#)

Public: Oui Mis en avant: Non

Nombre total de contenus

313



Dublin Core

Titre Les Infiniment petits

Sujet [SHS:LT] Humanities and Social Sciences/Literature

Créateur Lechevreil, Nadège

Source ANR-13-FRAL-0013

Éditeur <http://www.nakala.fr/account/11280/f1401838>

Droits Public Domain

Format TXT UTF-8

Type Text

Identifiant https://fr.wikisource.org/wiki/Les_Infiniment_petits

11280/d0d6667f

2016

parution 2016-09-09T12:38:18.887-02:00

Date de modification 2016-09-09T12:38:18.887-02:00

Licence <https://creativecommons.org/publicdomain/mark/1.0/>

Contenu suivant

Mé

Voir la pa

Sup

Public: Non M

Collection

nakala.biogra

Métadonné

Aucun fichier n
ajouté à ce Con
fichier.

Formats de s

- METS
- atom
- dc-rdf

En conclusion

Un corpus :

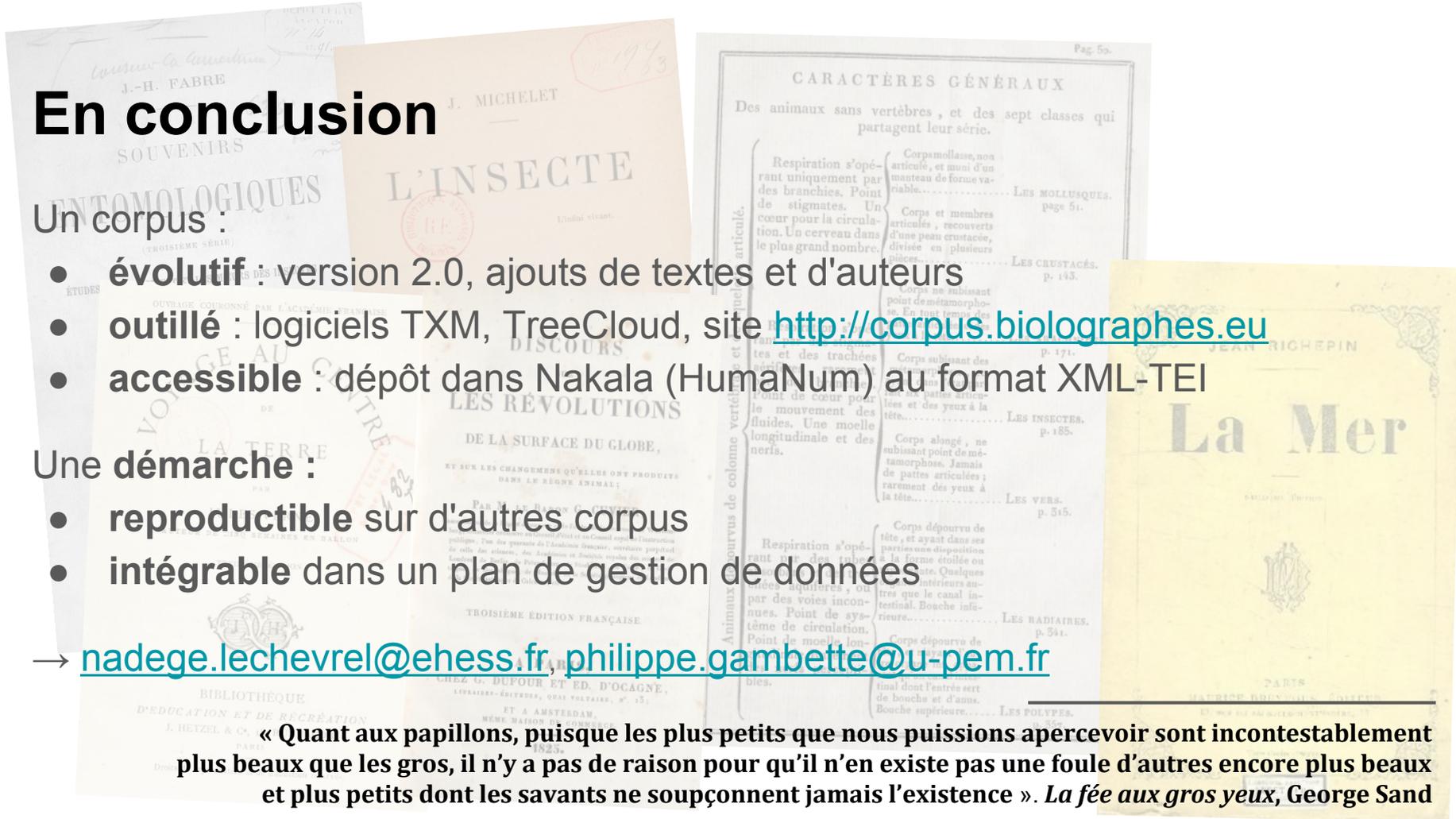
- **évolutif** : version 2.0, ajouts de textes et d'auteurs
- **outillé** : logiciels TXM, TreeCloud, site <http://corpus.biographes.eu>
- **accessible** : dépôt dans Nakala (HumaNum) au format XML-TEI

Une démarche :

- **reproductible** sur d'autres corpus
- **intégrable** dans un plan de gestion de données

→ nadege.lechevrel@ehess.fr, philippe.gambette@u-pem.fr

« Quant aux papillons, puisque les plus petits que nous puissions apercevoir sont incontestablement plus beaux que les gros, il n'y a pas de raison pour qu'il n'en existe pas une foule d'autres encore plus beaux et plus petits dont les savants ne soupçonnent jamais l'existence ». *La fée aux gros yeux*, George Sand



Références

- Nadège Lechevrel (2015) [Réception et vulgarisation des savoirs biologiques dans le corpus](#), billet du carnet *Biographes*
- Nadège Lechevrel & Philippe Gambette (2016) [Une approche textométrique pour étudier la transmission des savoirs biologiques au XIXe siècle](#), *Nouvelles perspectives en sciences sociales* 12(1), 221-253
- Nadège Lechevrel, Bénédicte Percheron & Gisèle Séginger (2016) Le polype. Formes et savoirs, série de billets du carnet *Biographes* :
 - [1. Le champ sémantique du polype : expérimentations numériques](#)
 - [3. Polype dans L'Encyclopédie de Diderot et d'Alembert](#)
 - [5. Le polype littéraire](#)
- Nadège Lechevrel, Philippe Gambette (2016), [Des savoirs partagés](#), billet du carnet *Biographes*
- Philippe Gambette, Nadège Lechevrel (2016), [Explorer les réseaux de citations du corpus Biographes](#), billet du carnet *Biographes*