Corpus Biolographes version 1.0 : construire, analyser, cartographier

26 janvier 2017
FMSH

Nadège Lechevrel
Philippe Gambette

Plan

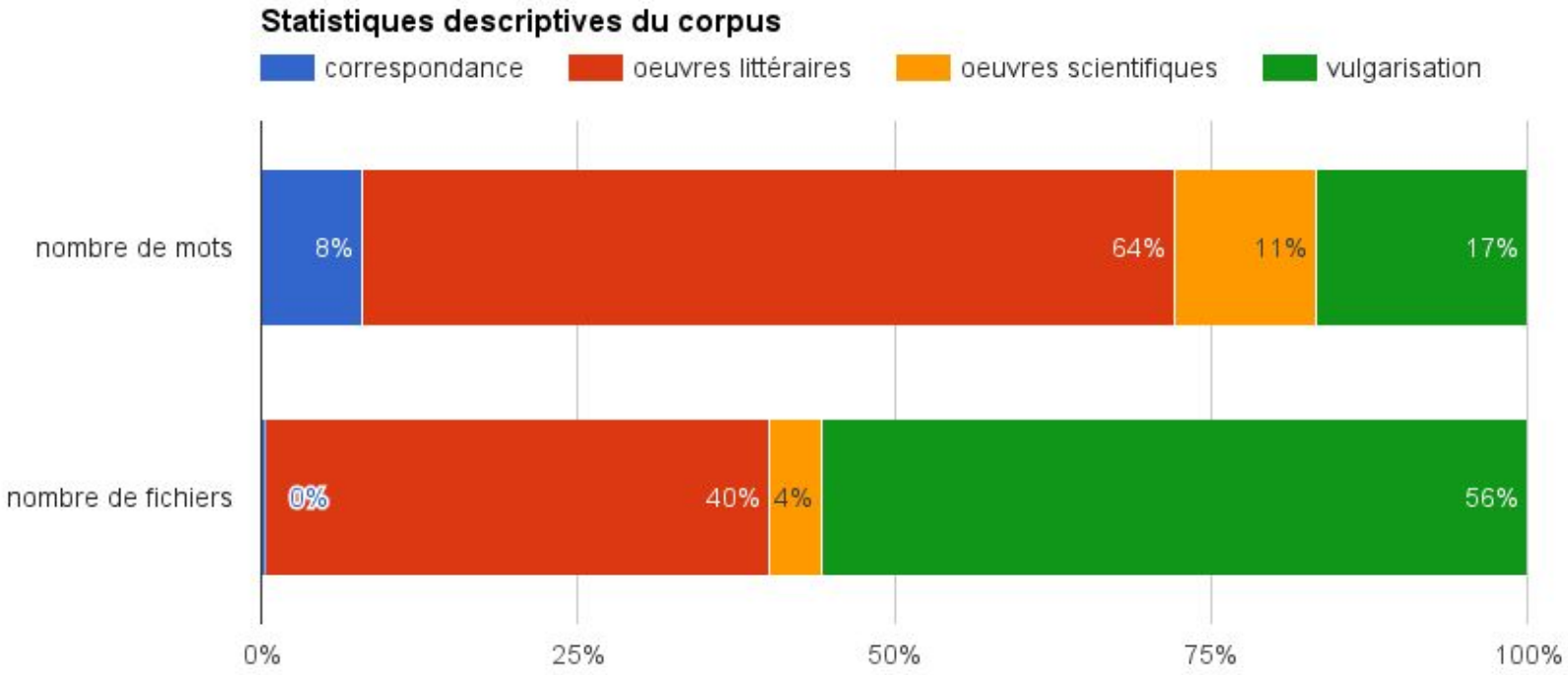
- Présentation du corpus Biolographes version 1.0
- Construire le corpus (Google Drive, PHP+MySQL)
- Analyser le corpus (TXM, TreeCloud, CFP)
- Cartographier le corpus (Palladio, auteurs citants/cités)
- Archiver le corpus (HumaNum)

Présentation du corpus Biolographes 1.0

Histoire naturelle, Papillons (1854), Pellerin

ark:/12148/btv1b6938165v

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

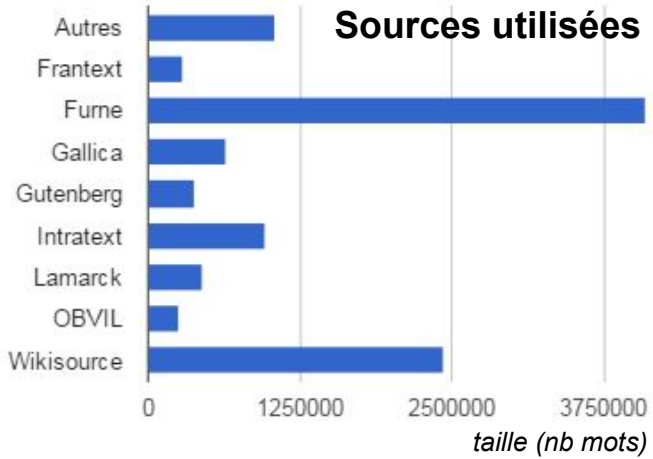
Parapluie pour la récolte des papillons.

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)



Construire le corpus : démarche suivie

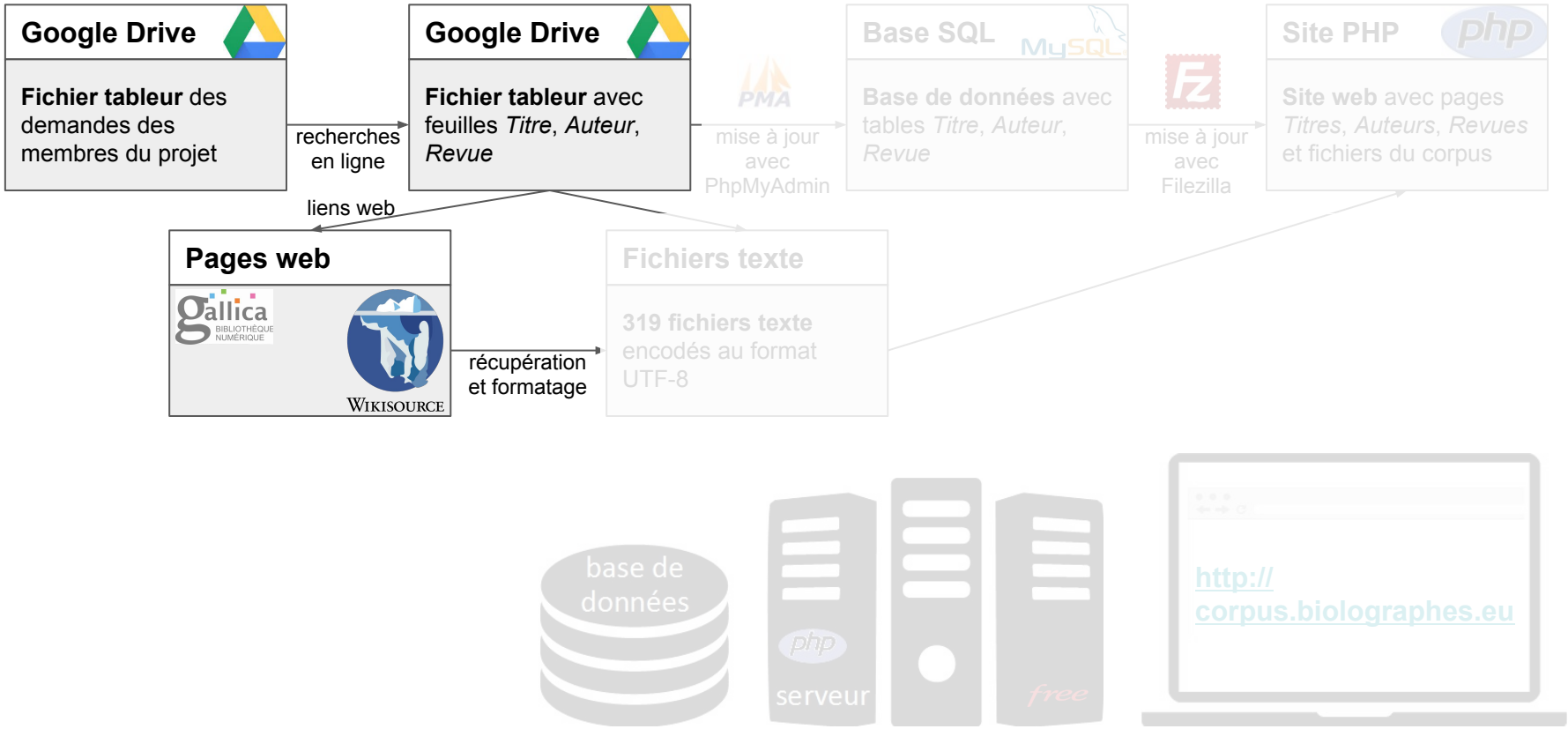
- Un fichier partagé pour réunir les demandes

Auteur	Titre	Revue	Numero	Tome	Annee	Pages	URL	Droits (domaine public)	Sujet (mots-cles)	Notes	Etat	Membre du projet à l'ori
Flaubert	Bouvard et Pécuchet.				2008		http://flaubert.univ-rouen.fr/bouvard_et_pecuchet/chantier/roman_final/BPframe1.html	http://flaubert.univ-rouen.fr/bouvard_et_pecuchet/mentions_legales.php	%classification prédéfinie? à voir avec les RAMEAU	Éd. de Stéphanie Dord-Crouslé. GF-Flammarion, 2008	ajouté au corpus	Gisèle Seginger
Flaubert	Lettres de Flaubert (1830-1880), édition Conard, 1926-1930.				1830-80		http://flaubert.univ-rouen.fr/correspondance/conard/outils/annuels.html	Toute réutilisation à des fins privées, à des fins d'enseignement ou de recherche scientifique est autorisée sous réserve d'indiquer "Éd. Danielle Girard et Yvan Leclerc, Rouen, 2003"		Éd. Danielle Girard et Yvan	ajouté au corpus	Gisèle Seginger
Fabre	Souvenirs entomologiques (livre 1-10).				1879-1907		http://www.intratext.com/Catalogo/Autori/AUT1925.HTM			Textes intégraux des *Souvenirs* également sur e-fabre, mais l'édition et la série ne sont pas indiqués (http://www.e-fabre.com/biographie/souvenirs_sommaire.htm). Idem sur IntraText	ajouté au corpus	Gisèle Seginger
Fabre	Texte scientifique d'entomologie (lesquels?										en cours	Gisèle Seginger
Henri Milne Edwards	Recherches anatomiques, physiolc				1838		http://biodiversitylibrary.org/item/108103#page/9/mode/1up			OCR non corrigé	en cours	Nadège Lechevrel
Henri Milne Edwards	Histoire naturelle des coralliaires :				1857-1860		http://biodiversitylibrary.org/item/44590#page/9/mode/1up			OCR non corrigé	en cours	Nadège Lechevrel
Flaubert	Salammbô				1910		https://fr.wikisource.org/wiki/Salammb%C3%B4	domaine public			ajouté au corpus	Gisèle Séginger
Michelet	Journal de Michelet										en cours	Gisèle Séginger
Darwin	L'origine des espèces				1862		https://fr.wikisource.org/wiki/L%E2%80%99Origine_des_esp%C3%A8ces/Notice				ajouté au corpus	tous
Boussenard	Les Secrets de Monsieur Synthèse				1888		https://beq.ebooksgratuits.com/vents/Boussenard-Synthese.pdf	domaine public			ajouté au corpus	Bénédicte Percheron
Mirbeau Octave	Jardin des supplices				1899		https://fr.wikisource.org/wiki/Le_Jardin_des_supplices	domaine public		Bibliothèque Charpentier, 1911	ajouté au corpus	Bénédicte Percheron
Maupassant	Contes et nouvelles						http://maupassant.free.fr/contes1.htm	domaine public		propre	en cours	Bénédicte Percheron
Lyell	Charles											

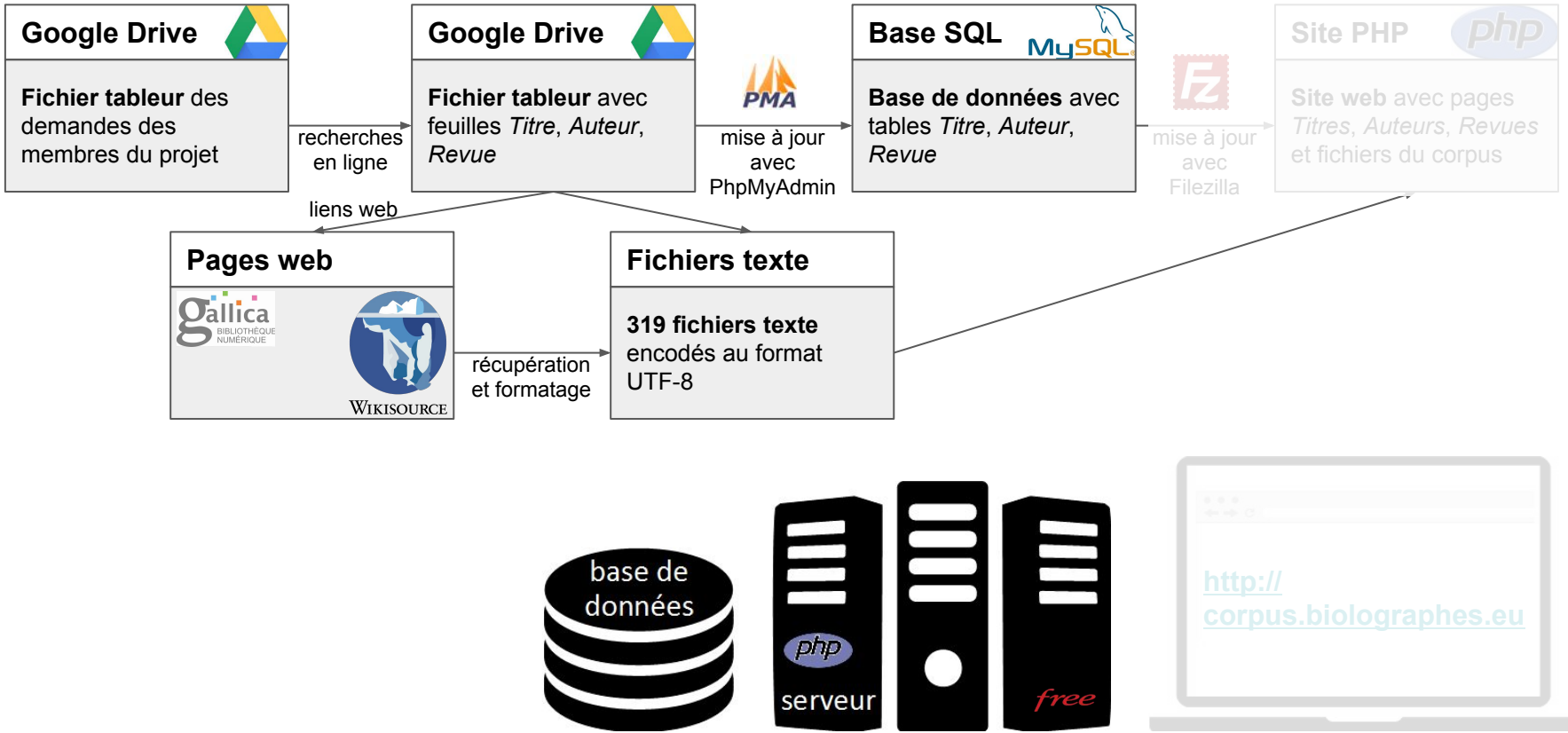
- Corrections manuelles de certains textes
- Contributions à WikiSource pour certains textes corrigés

Étapes de construction du corpus

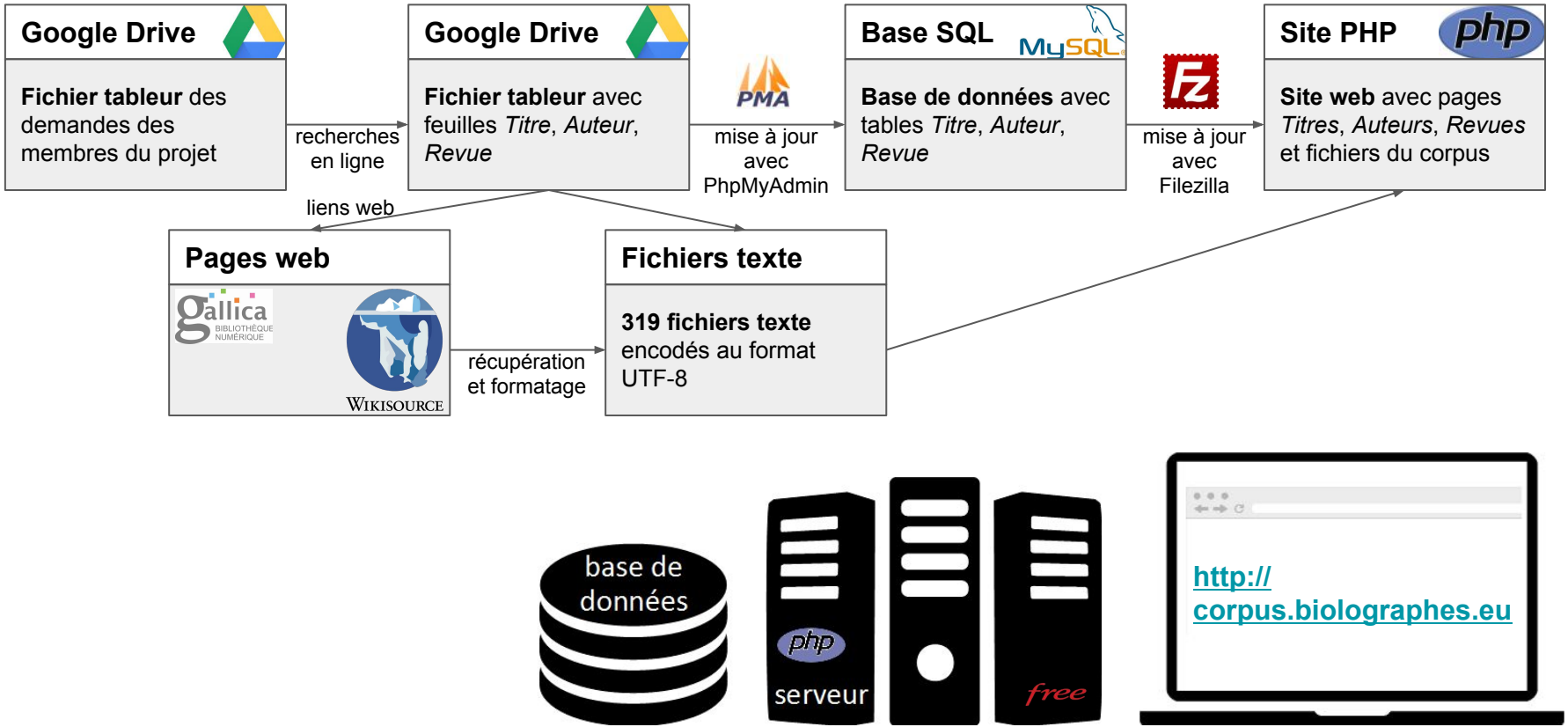
Étapes de construction du corpus



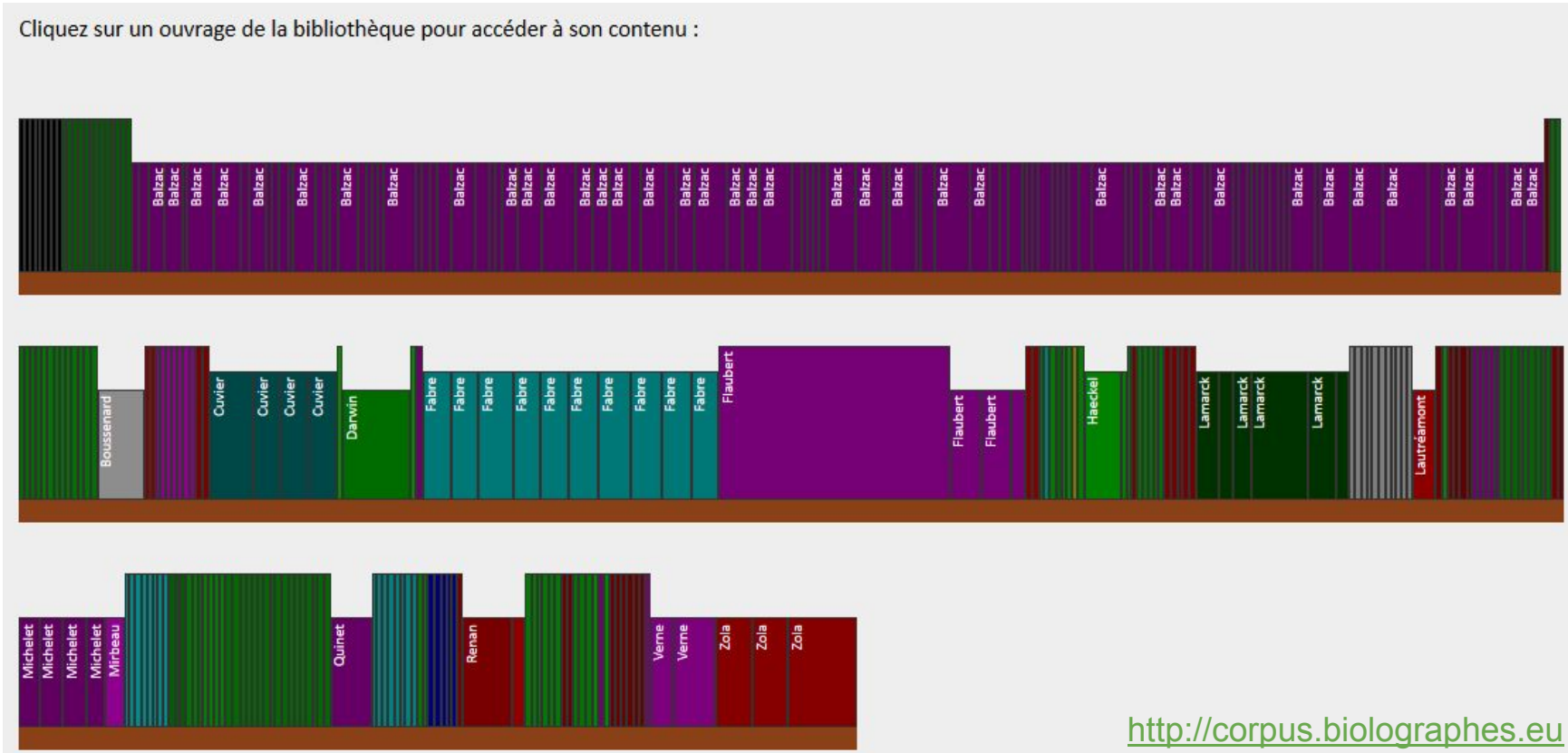
Étapes de construction du corpus



Étapes de construction du corpus



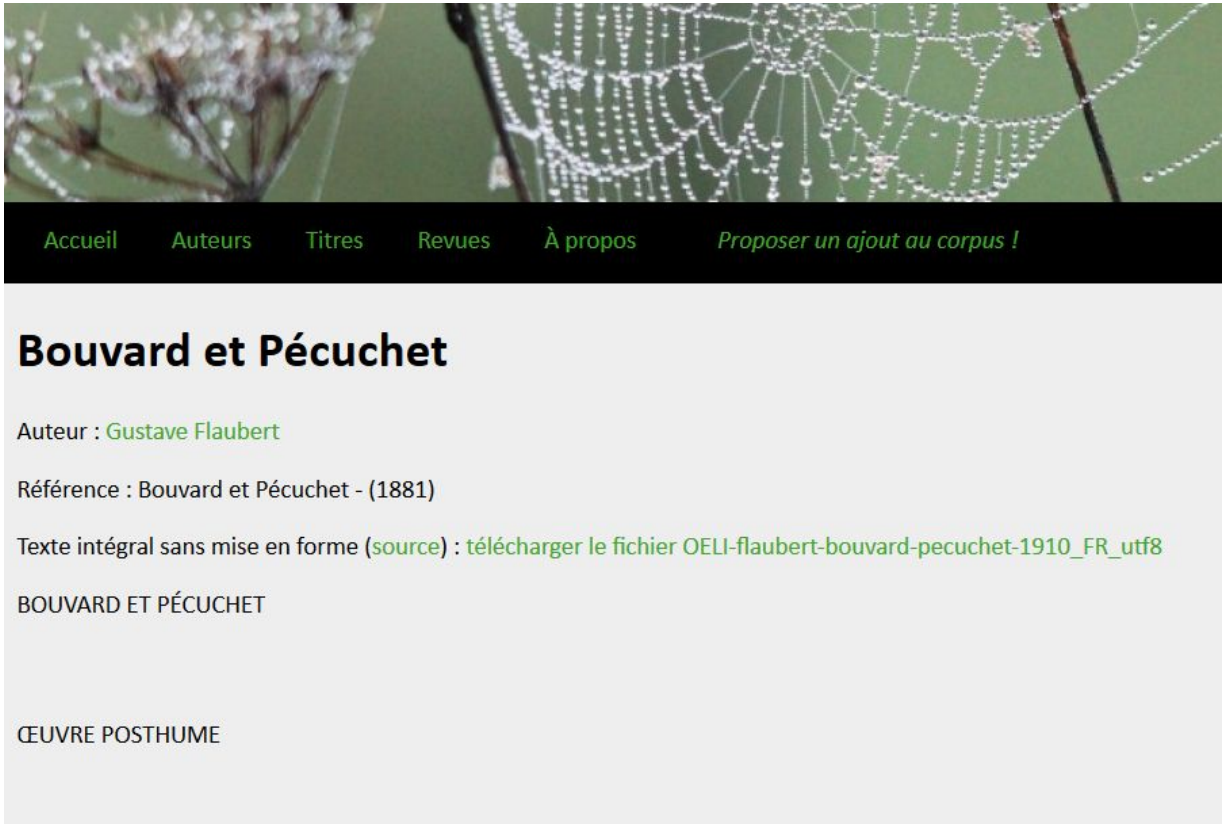
Parcours du site du corpus Biographes



http://corpus.biolographes.eu

Parcours du site du corpus Biolographes

Choix d'un ouvrage



http://corpus.biolographes.eu

Parcours du site du corpus Biolographes

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

Bouvard et Pécuchet

Auteur : Gustave Flaubert

Référence : Bouvard et Pécuchet - (1881)

Texte intégral sans mise en forme (source) : télécharger le fichier OELI-flaub

BOUVARD ET PÉCUCHET

ŒUVRE 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)

http://corpus.biolographes.eu

Parcours du site du corpus Biolographes

Plus d'informations : data.bnf.fr

Gustave Flaubert (1821

Romancier et auteur dramatique

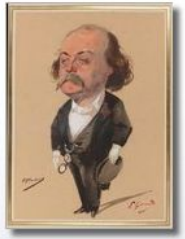
Plus d'informations

Œuvres présentes dans le corp

- 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) *(lituanien)*
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

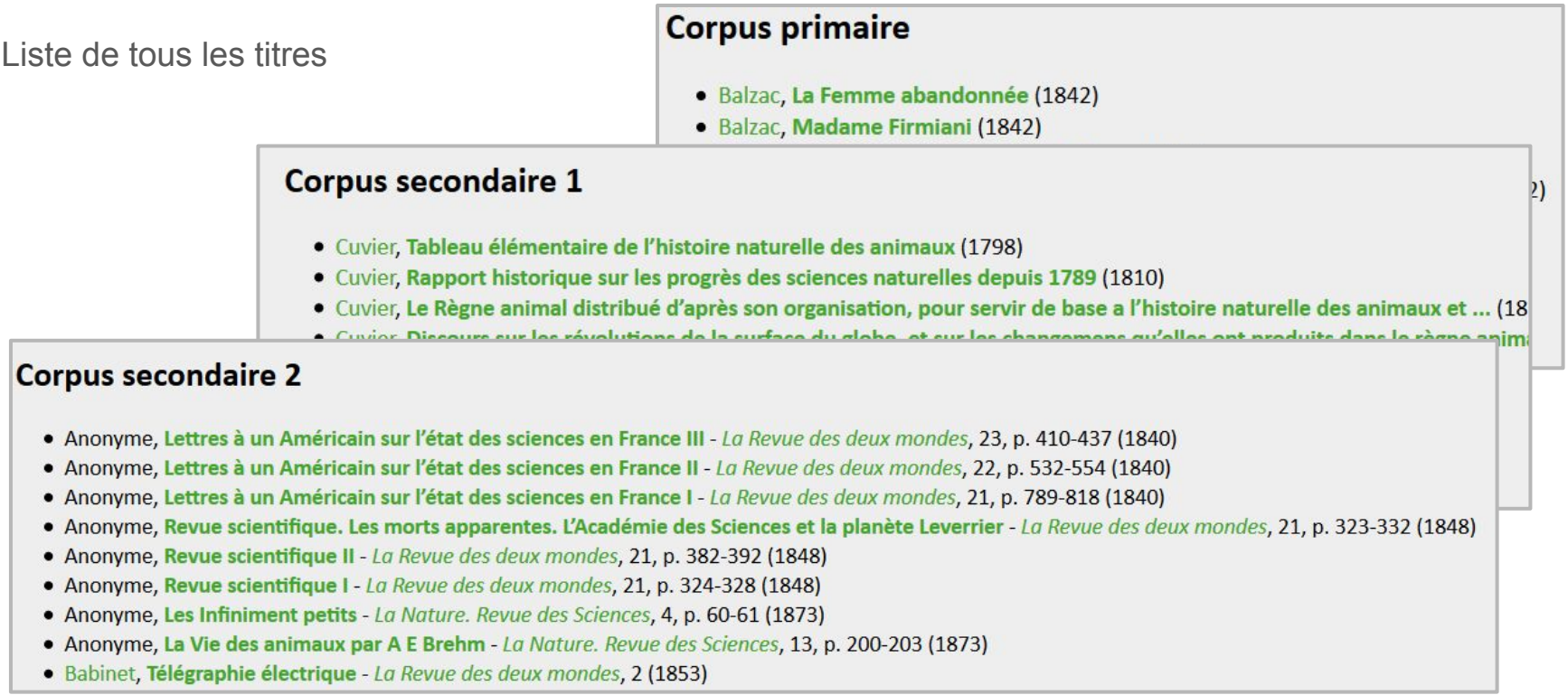
Ses activités

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

http://corpus.biolographes.eu

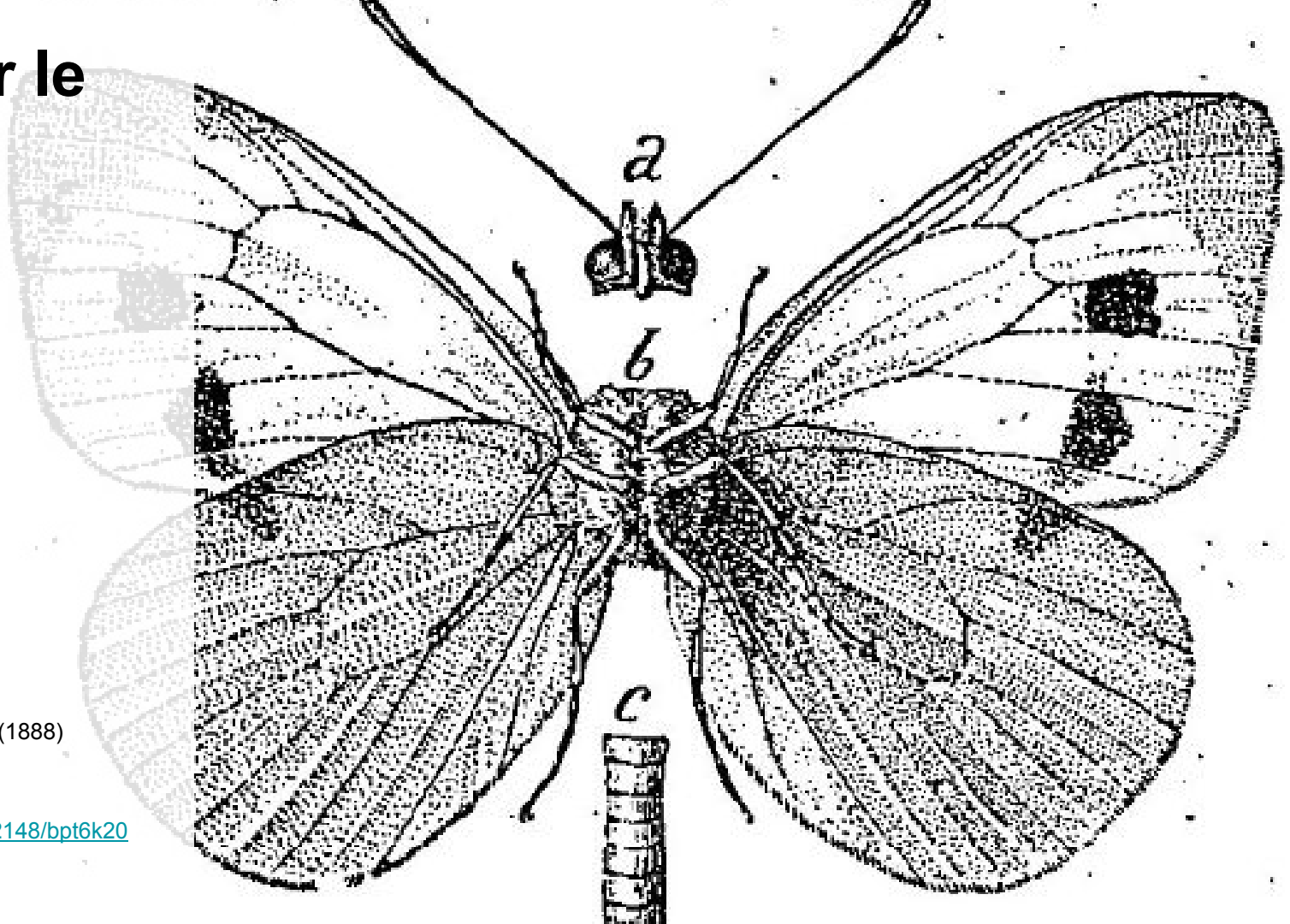
Parcours du site du corpus Biolographes

Liste de tous les titres



http://corpus.biolographes.eu

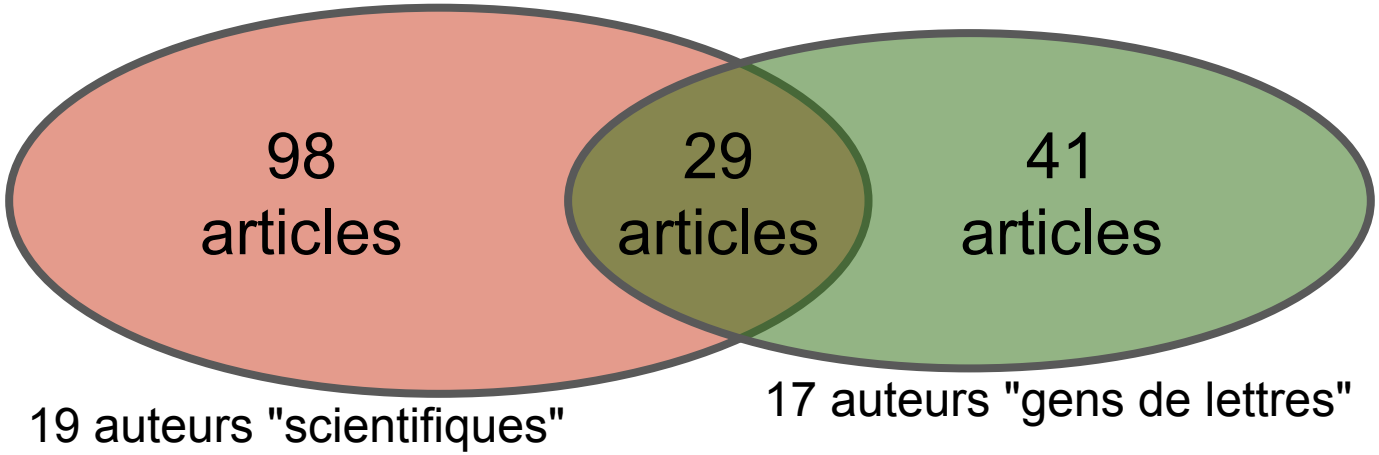
Analyser le corpus



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

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

Le sous-corpus *Revue des deux Mondes*



Informations quantitatives globales :

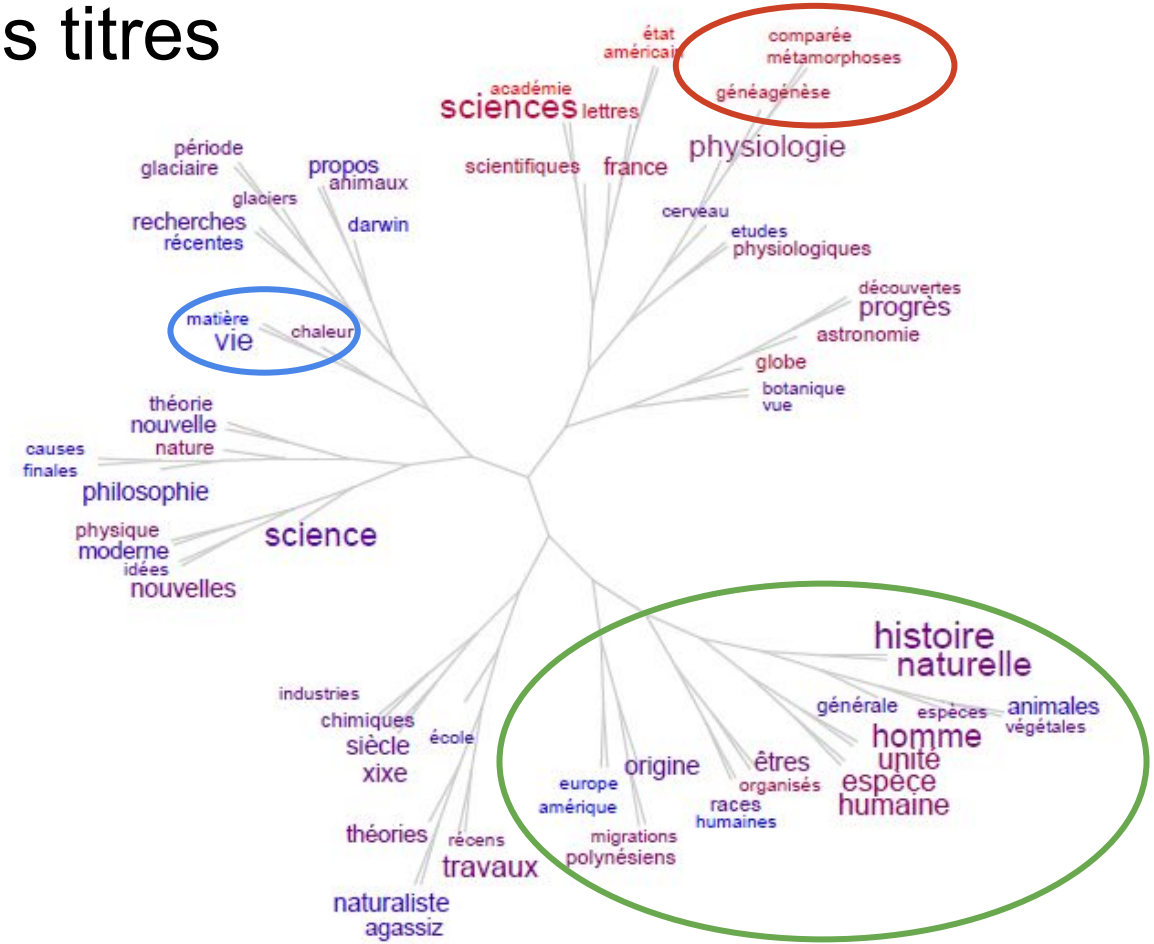
- 11 méga-octets
- environ 2 millions d'occurrences, 54 000 formes uniques
- 168 articles

Résumé visuel des titres

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

début de siècle

fin de siècle



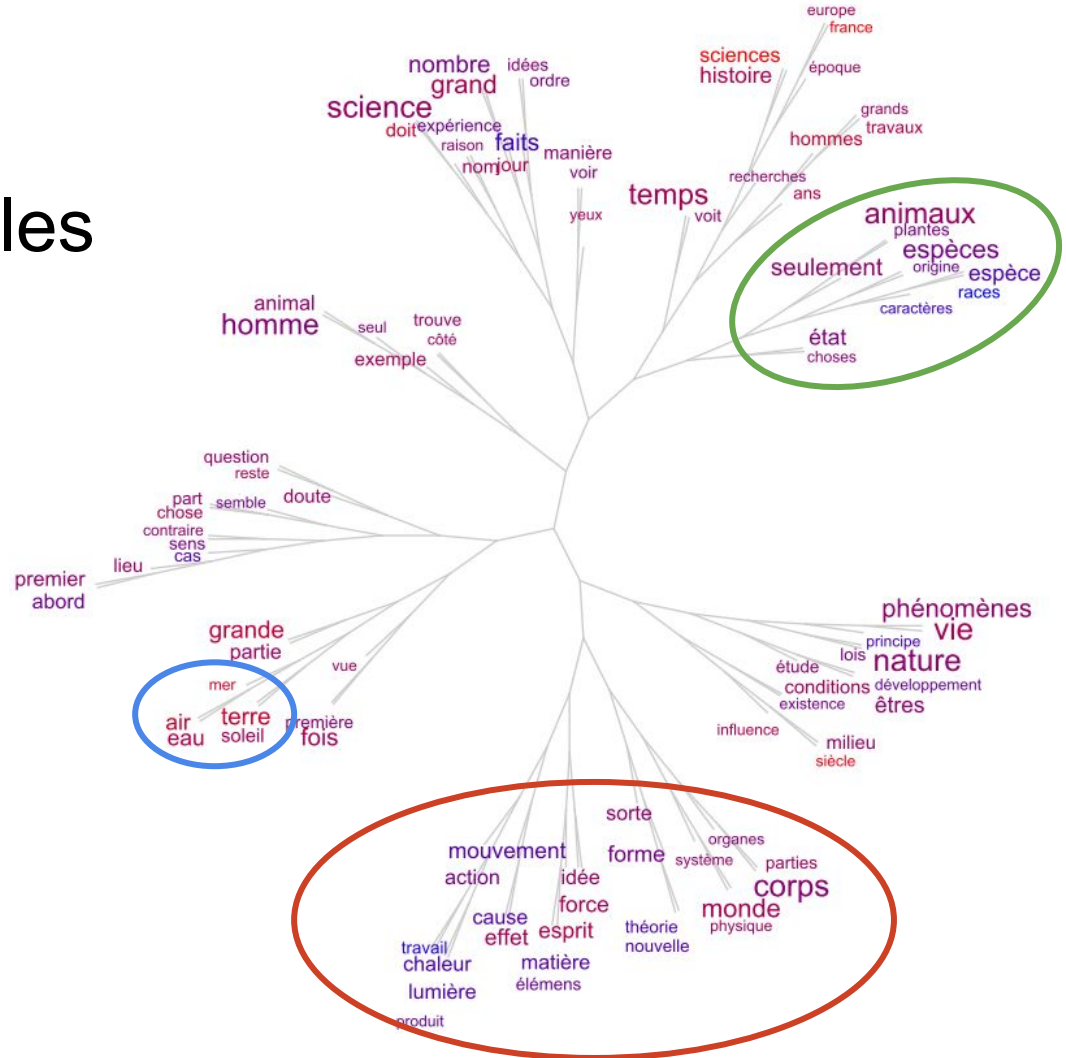
http://www.treecloud.org

Résumé visuel du contenu des articles

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

début de siècle

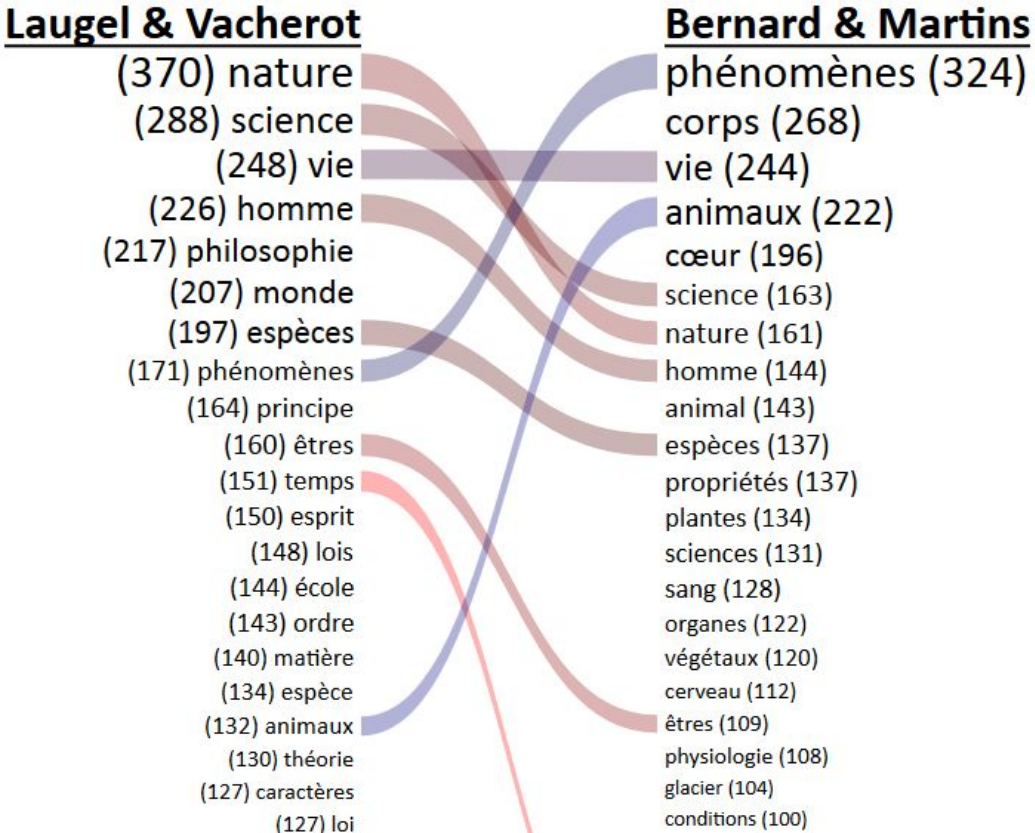
fin de siècle



http://www.treecloud.org

Vocabulaire des scientifiques et gens de lettres

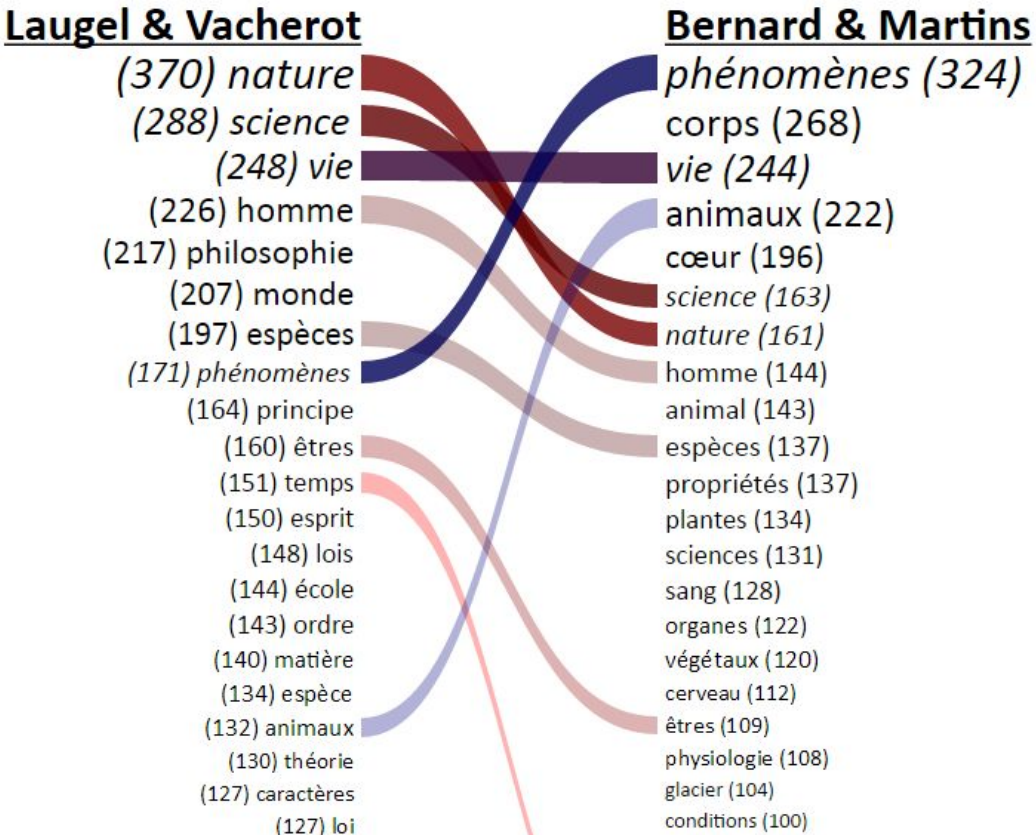
Chaînes de formes partagées



http://treecloud.univ-mlv.fr/cfp/index_fr.php

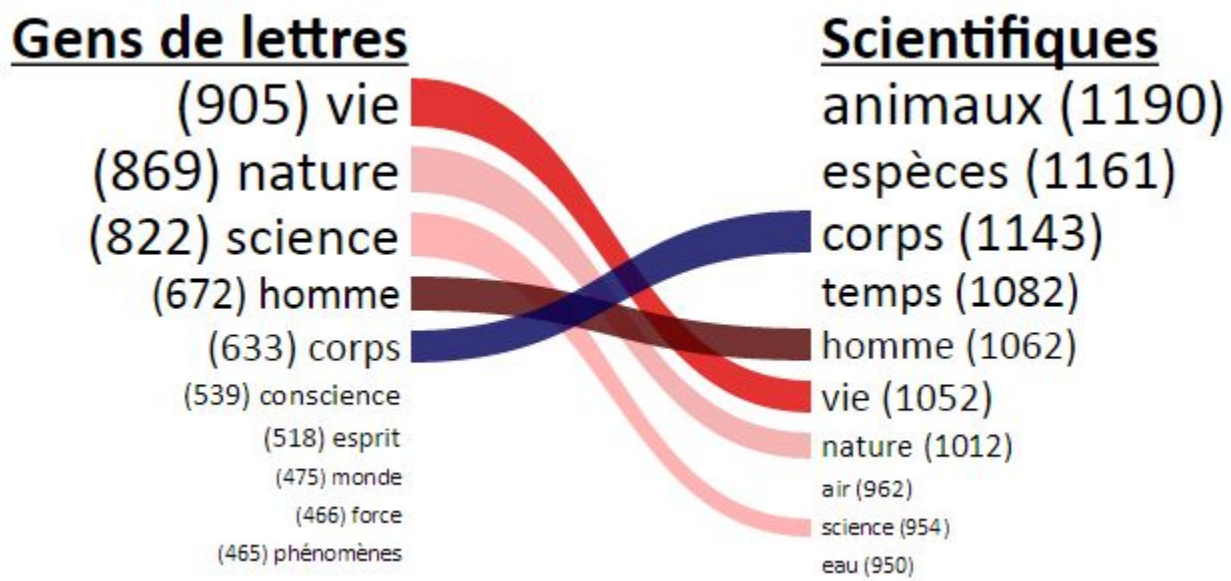
Vocabulaire des scientifiques et gens de lettres

Chaînes de formes partagées



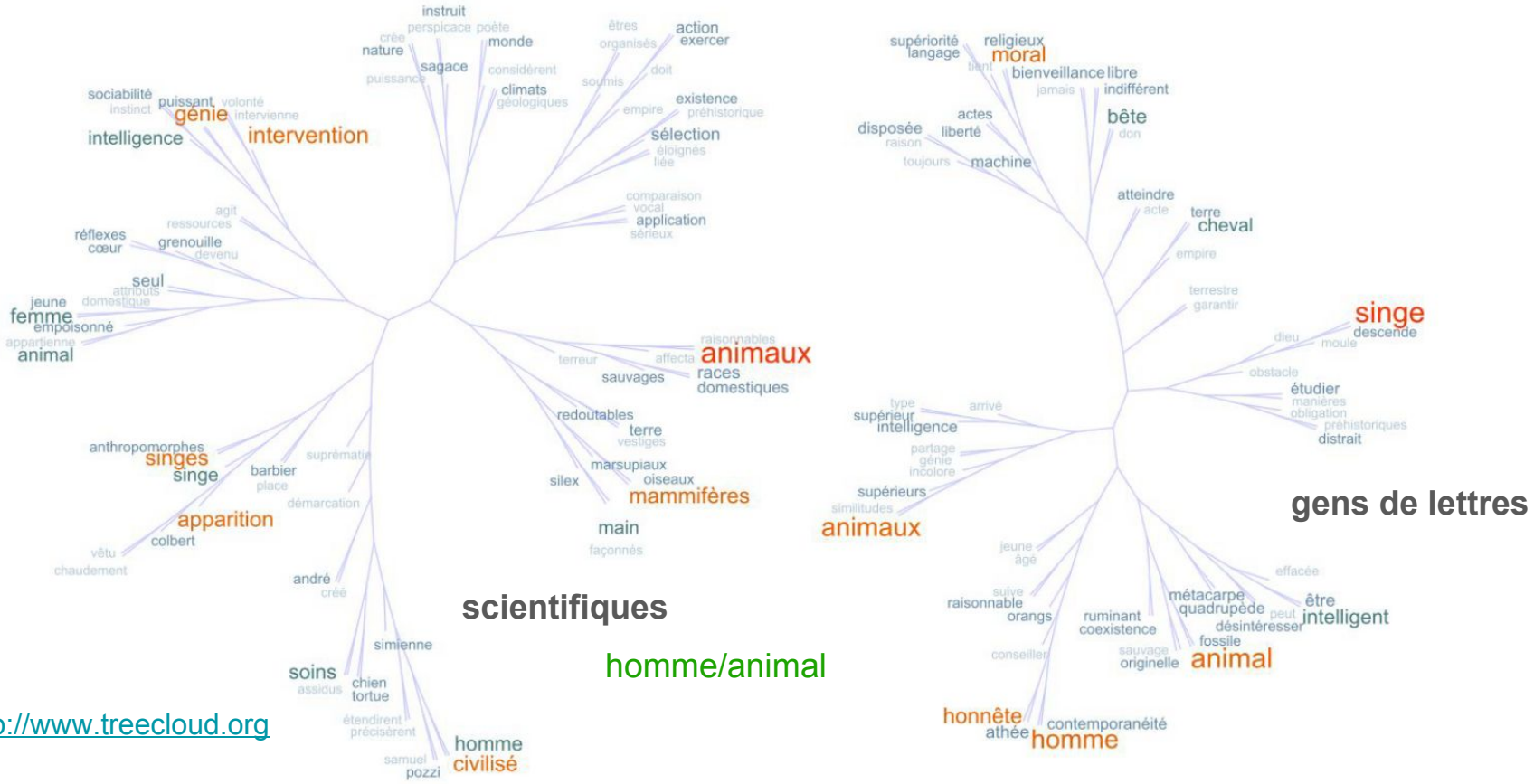
http://treecloud.univ-mlv.fr/cfp/index_fr.php

Vocabulaire des scientifiques et gens de lettres



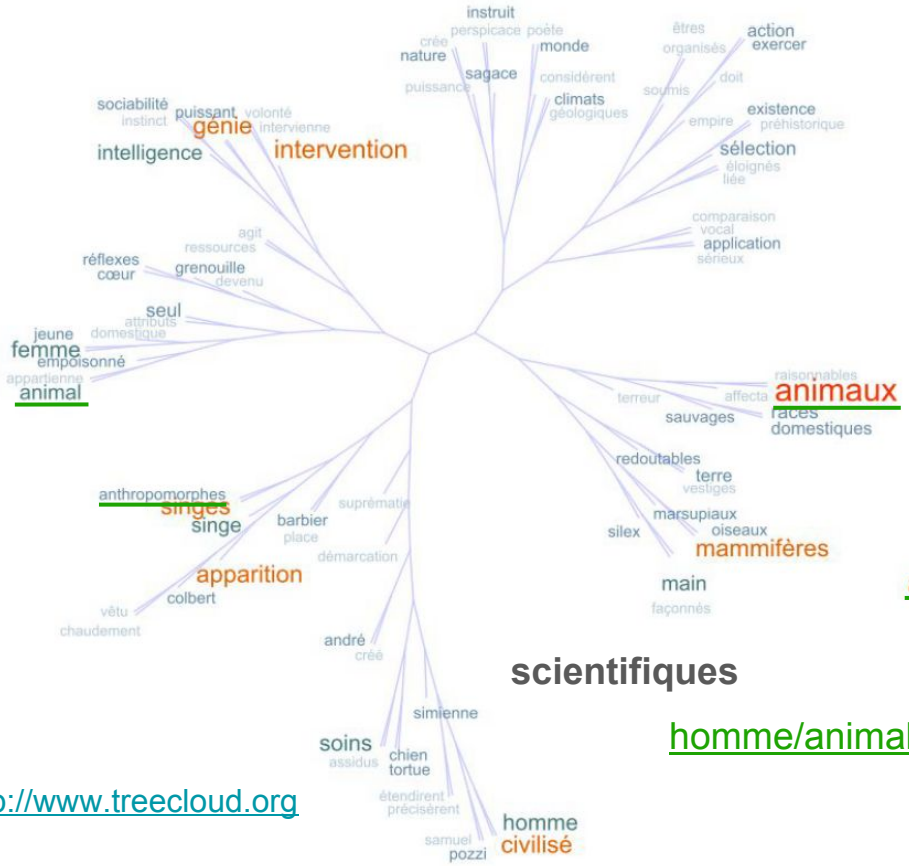
http://treecloud.univ-mlv.fr/cfp/index_fr.php

Nuages arborés des contextes de « homme »



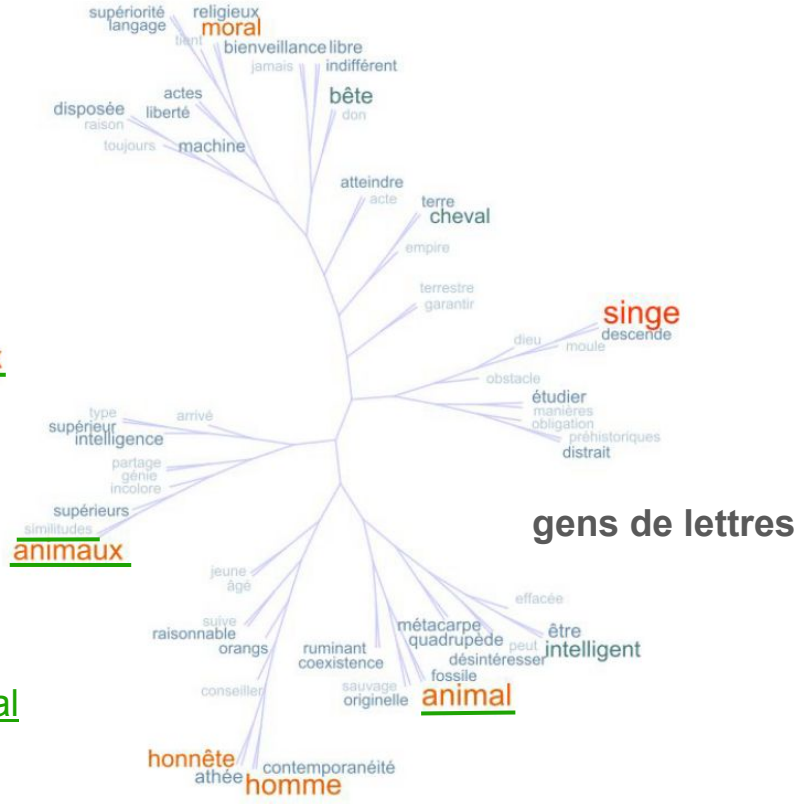
http://www.treecloud.org

Nuages arborés des contextes de « homme »



scientifiques

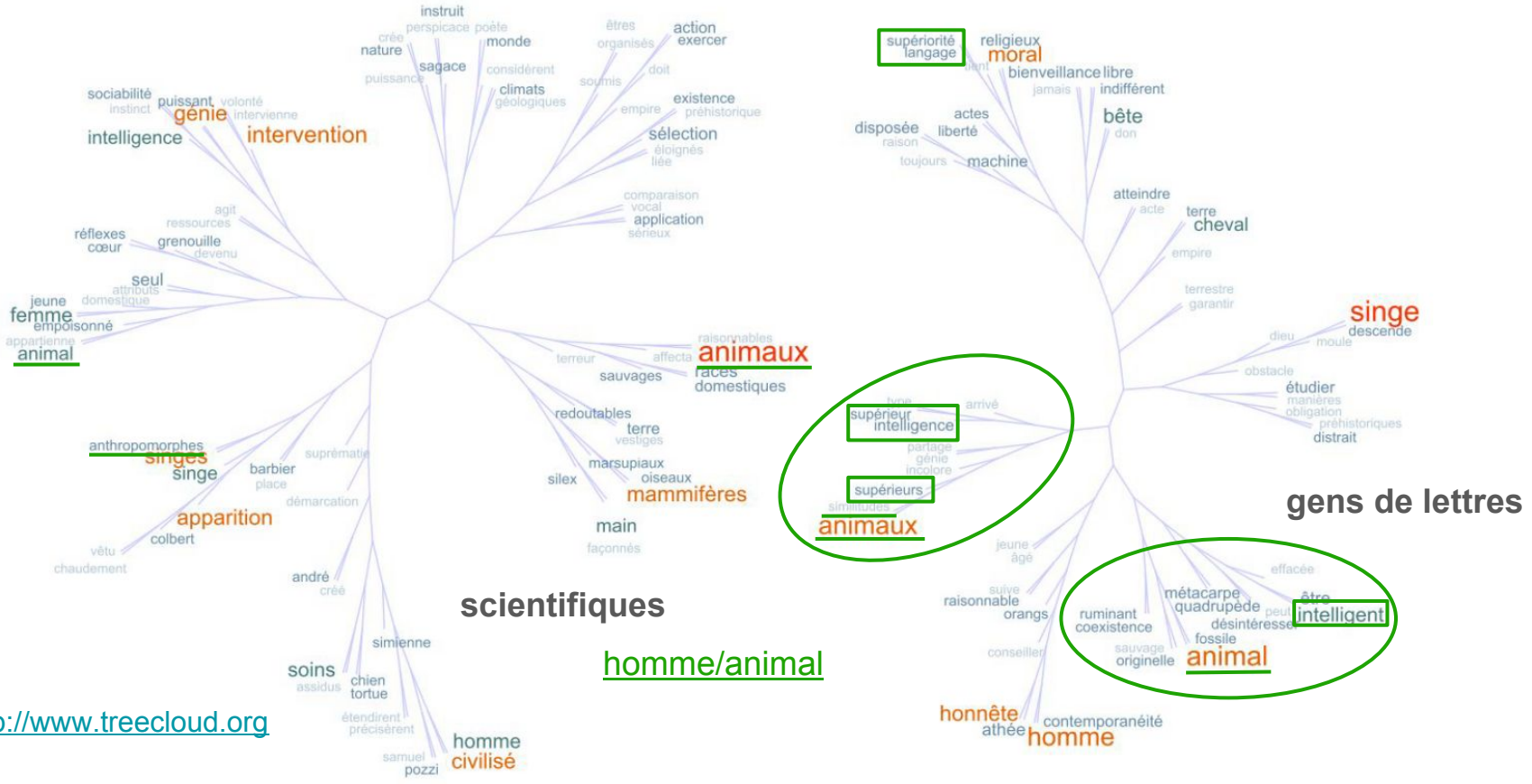
homme/animal



gens de lettres

http://www.treecloud.org

Nuages arborés des contextes de « homme »



scientifiques

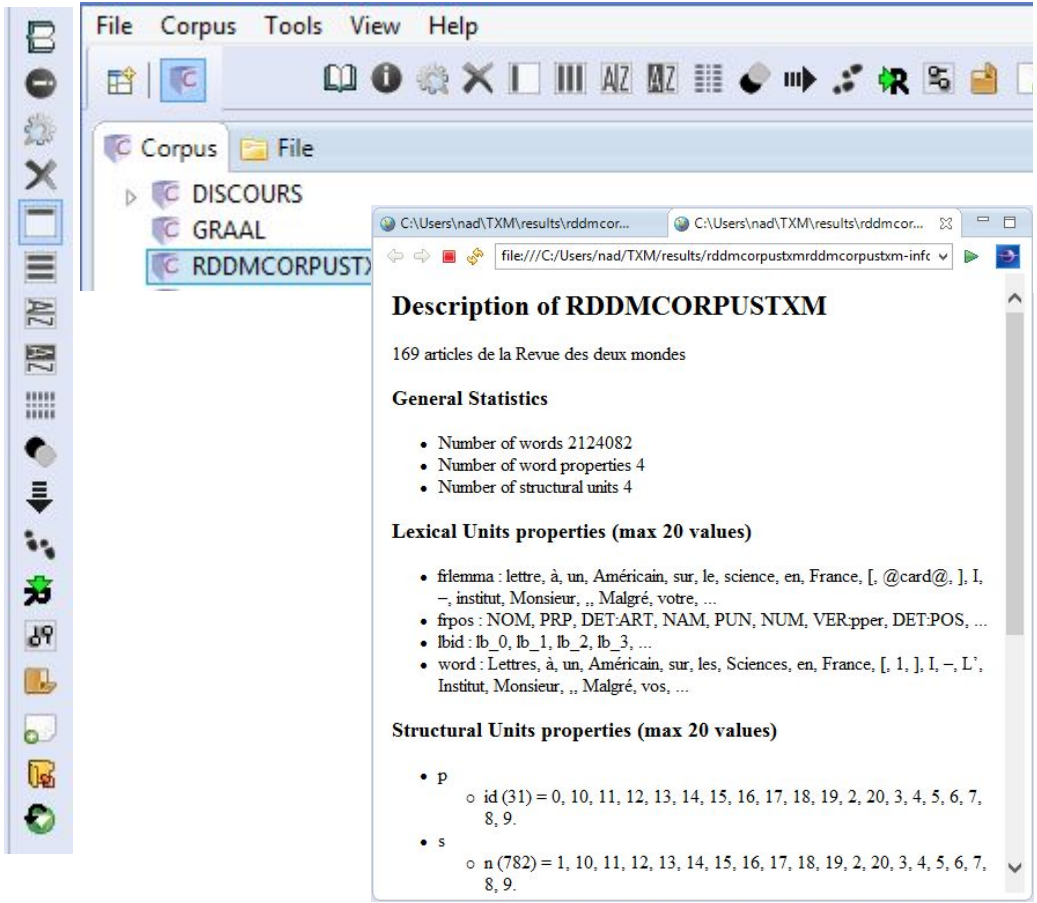
gens de lettres

homme/animal

http://www.treecloud.org

TXM

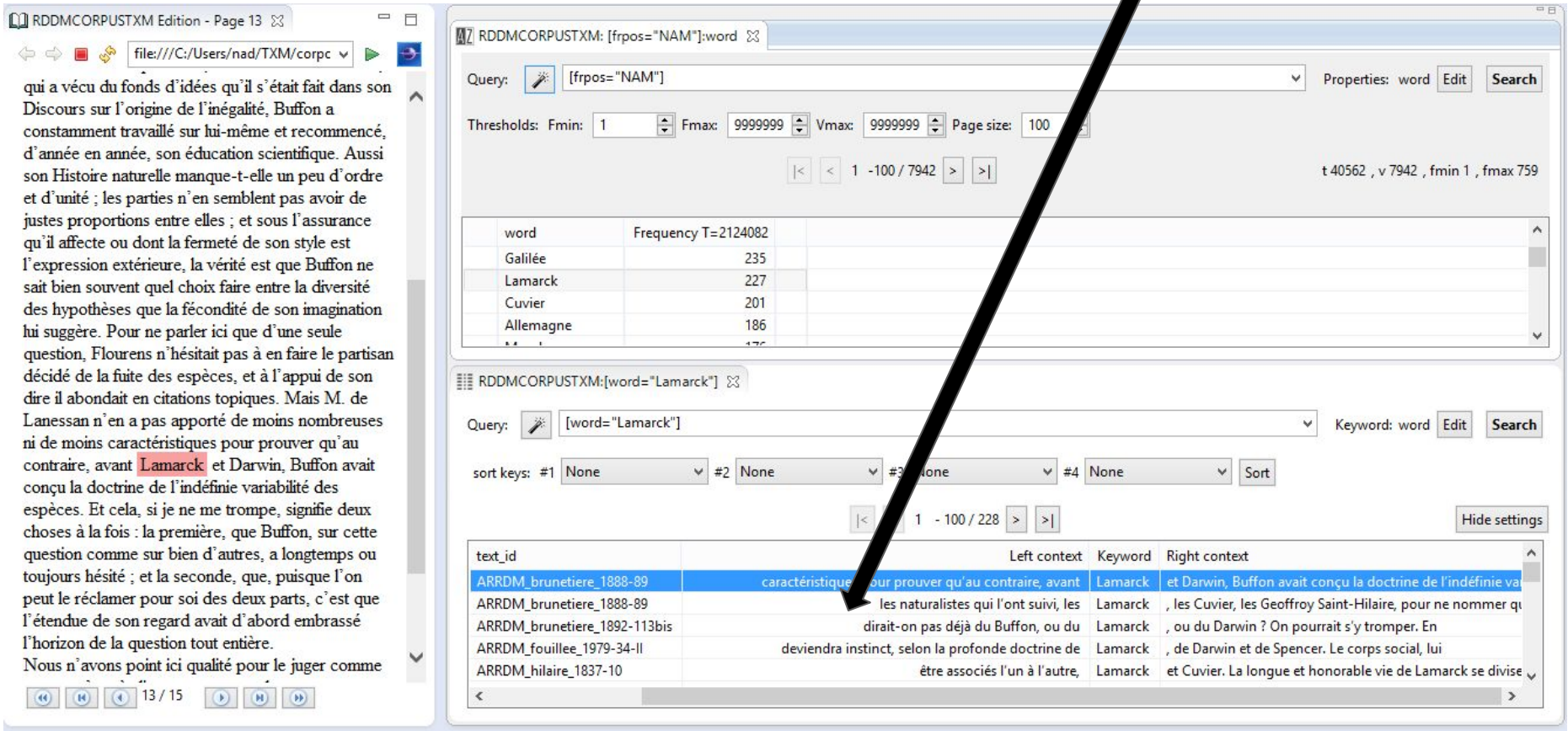
- Étiquetage des données
- Lemmatisation
- Édition des textes
- Partition du corpus
- Spécificités
- **Contextes**
- Étude du lexique
- Recherche de motifs linguistiques (CQL)
- Recherches contextuelles



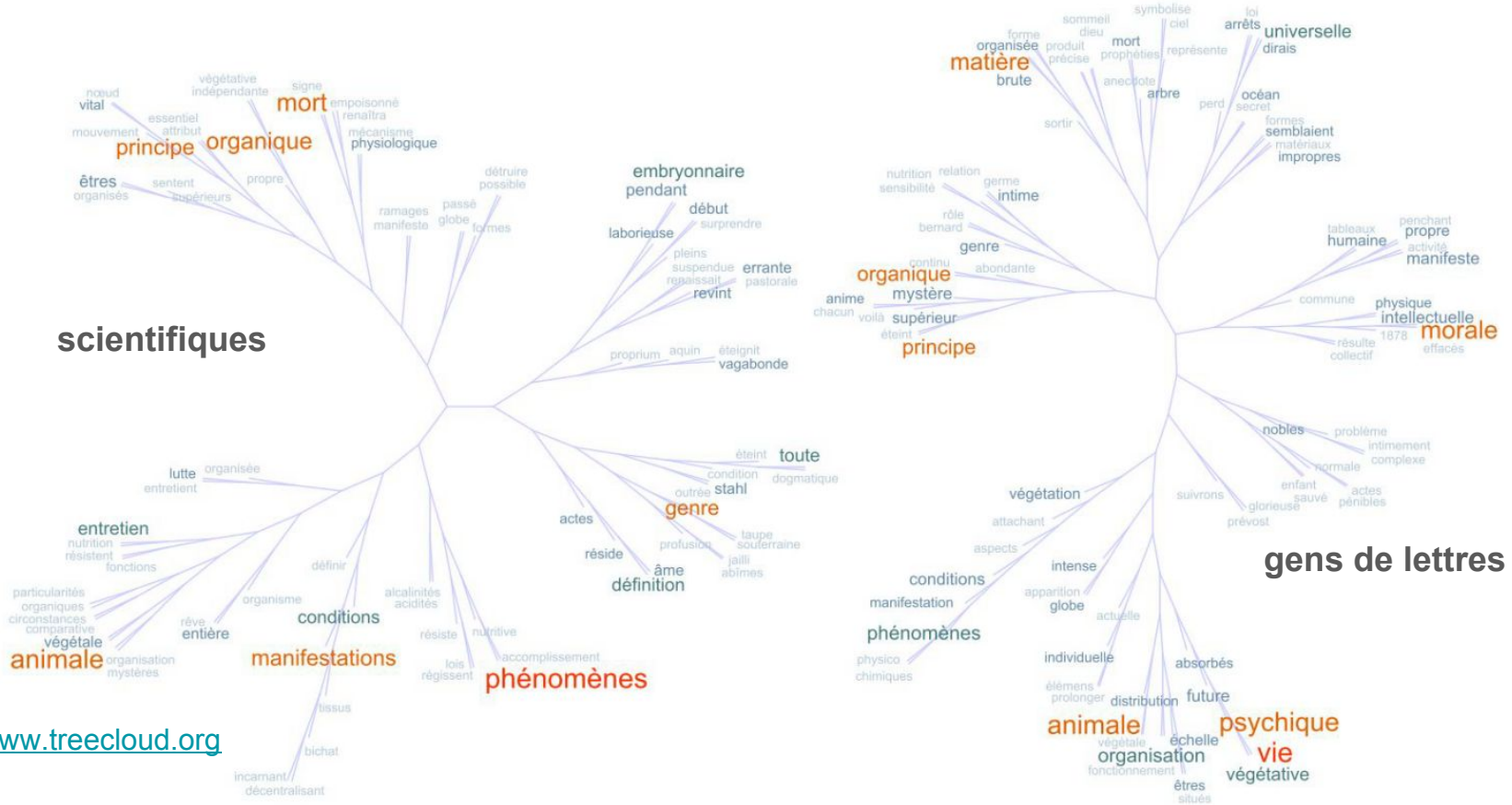
Accès au texte

Index et fréquence des termes

Concordancier

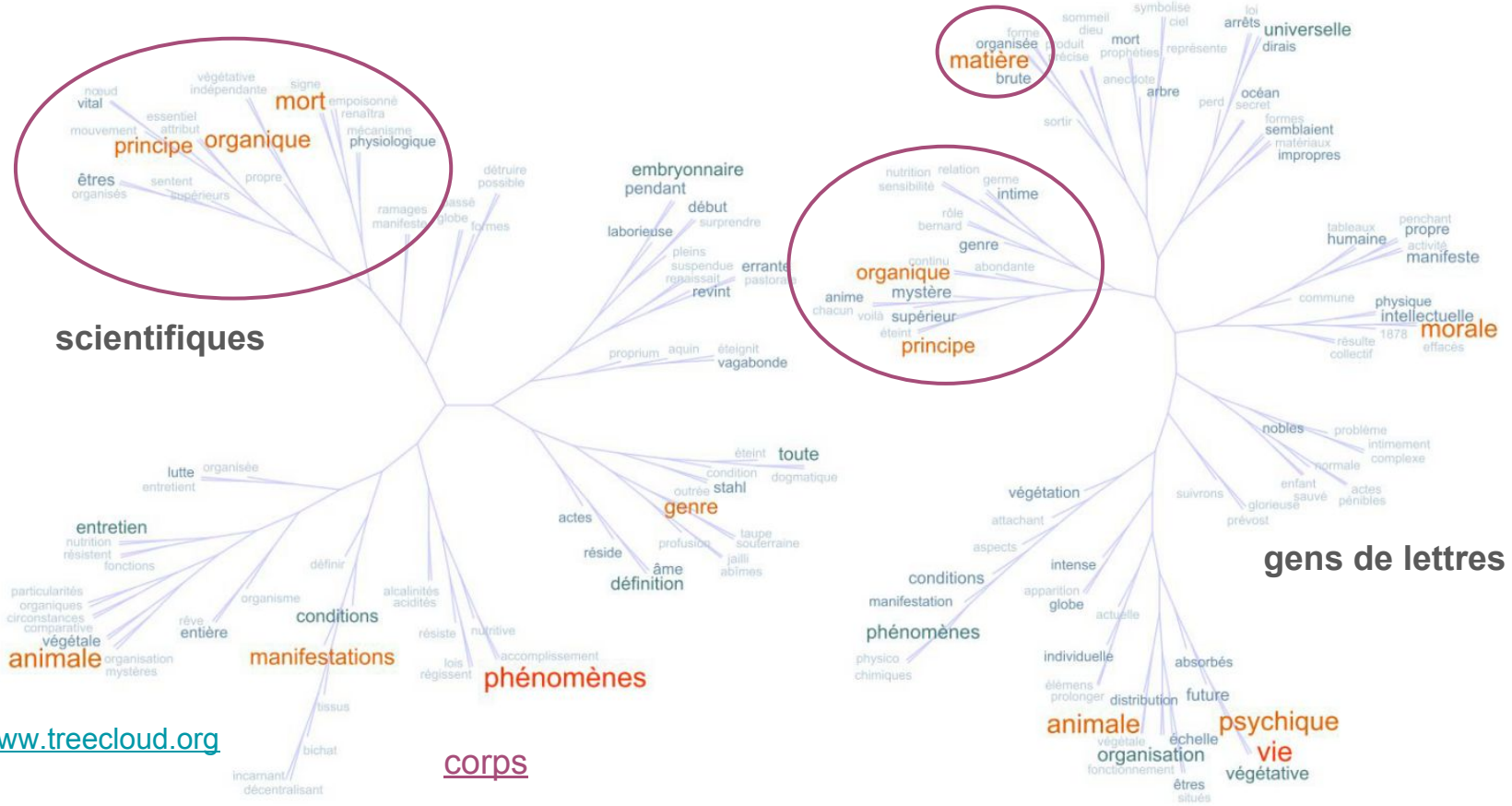


Nuages arborés des contextes de « vie »

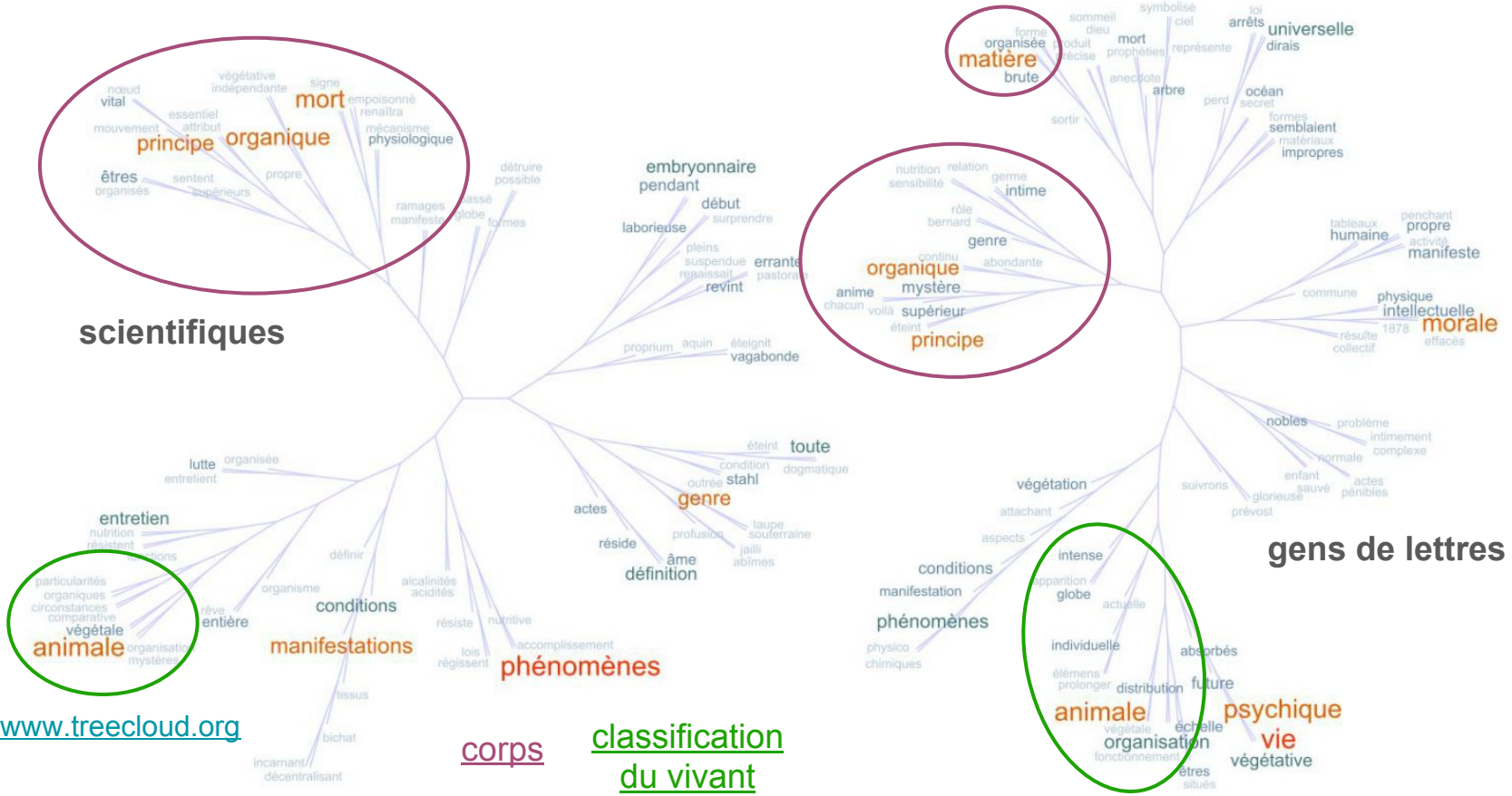


http://www.treecloud.org

Nuages arborés des contextes de « vie »



Nuages arborés des contextes de « vie »

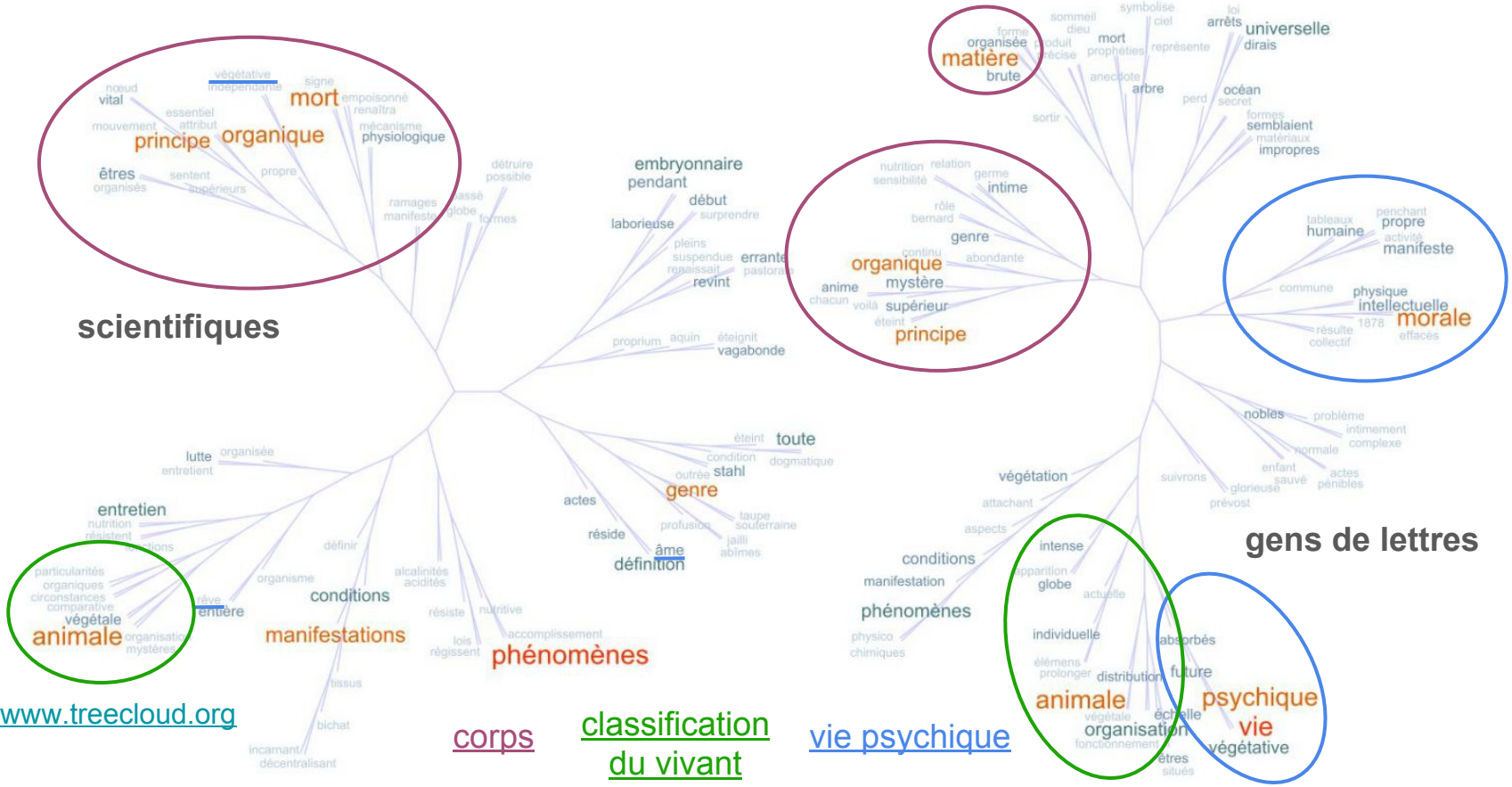


http://www.treecloud.org

corps

classification du vivant

Nuages arborés des contextes de « vie »



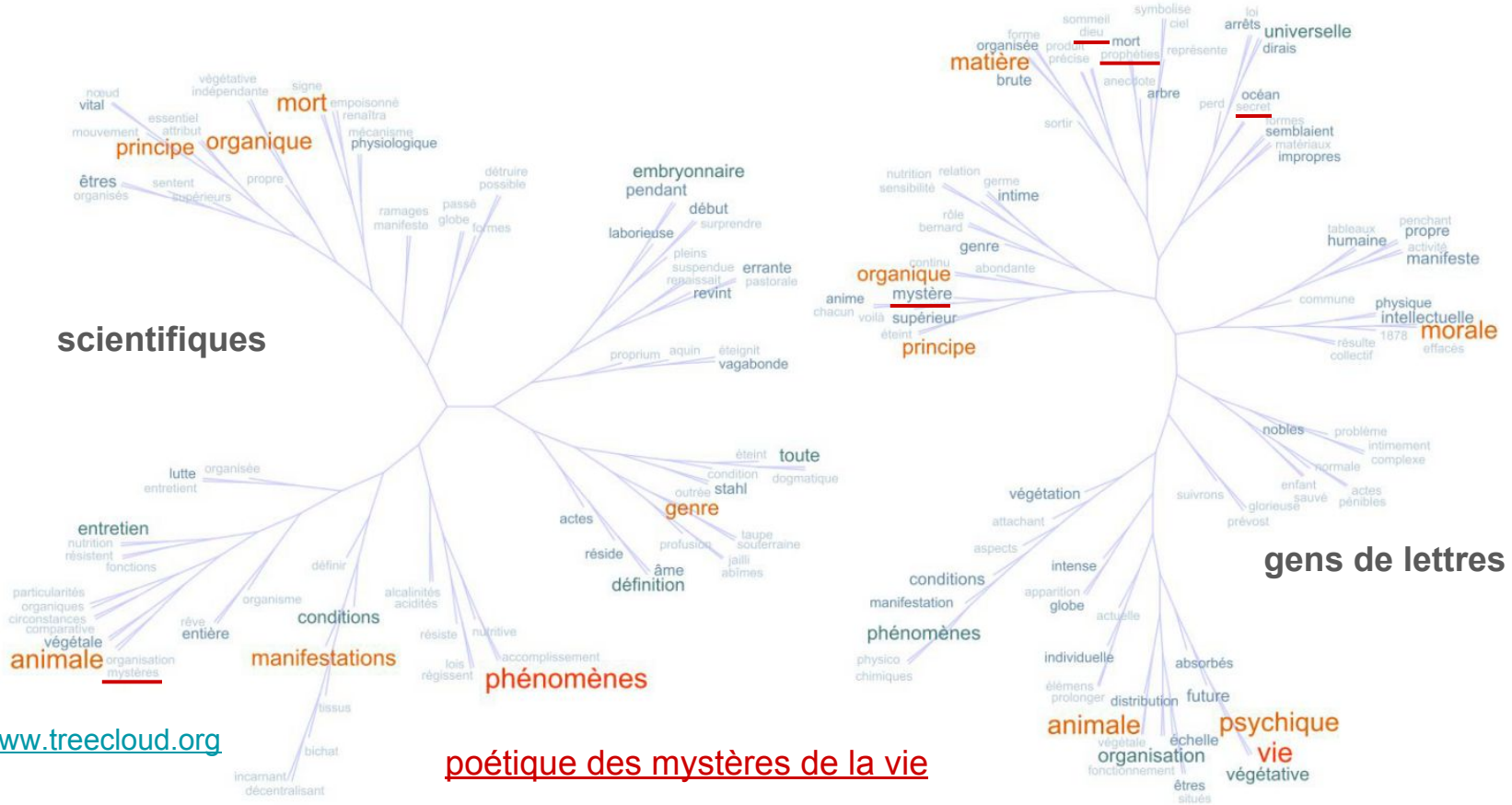
http://www.treecloud.org

corps

classification du vivant

vie psychique

Nuages arborés des contextes de « vie »



http://www.treecloud.org

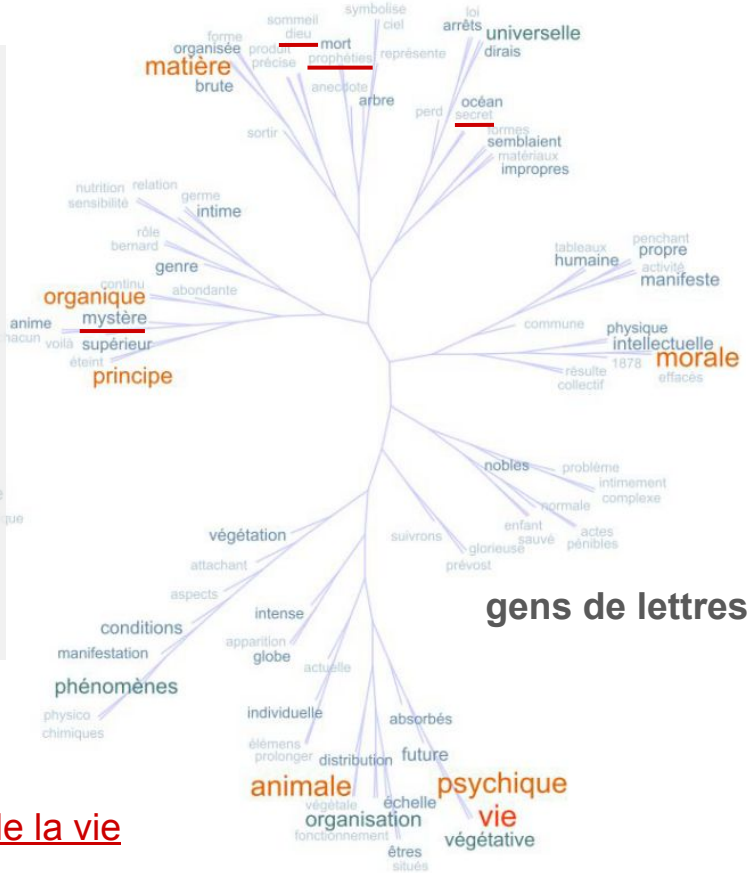
poétique des mystères de la vie

Nuages arborés des contextes de « vie »

« Le divorce entre le corps et l'âme était prononcé, et le grand Pan, le **dieu** de la **vie**, léguait à ses derniers adeptes la tâche de réhabiliter la matière. » George Sand

« N'est-il pas curieux de voir les prêtres français du XVIIe siècle croire comme les brahmanes à la puissance des fleurs lancées par l'arc du **dieu** de l'amour ? Image de la **vie** et de la fécondité dans notre monde, la fleur est dans l'autre la figure de l'immortalité. » Dora d'Istria

« la **vie**, selon ces médecins superstitieux, serait une sorte de **dieu** capricieux et de Protée menteur, échappant à toute prise, et avec lequel on ne peut lutter qu'au moyen de cette autre force, non moins aveugle et capricieuse, qu'ils appellent l'inspiration » Paul Janet



http://www.treecloud.org

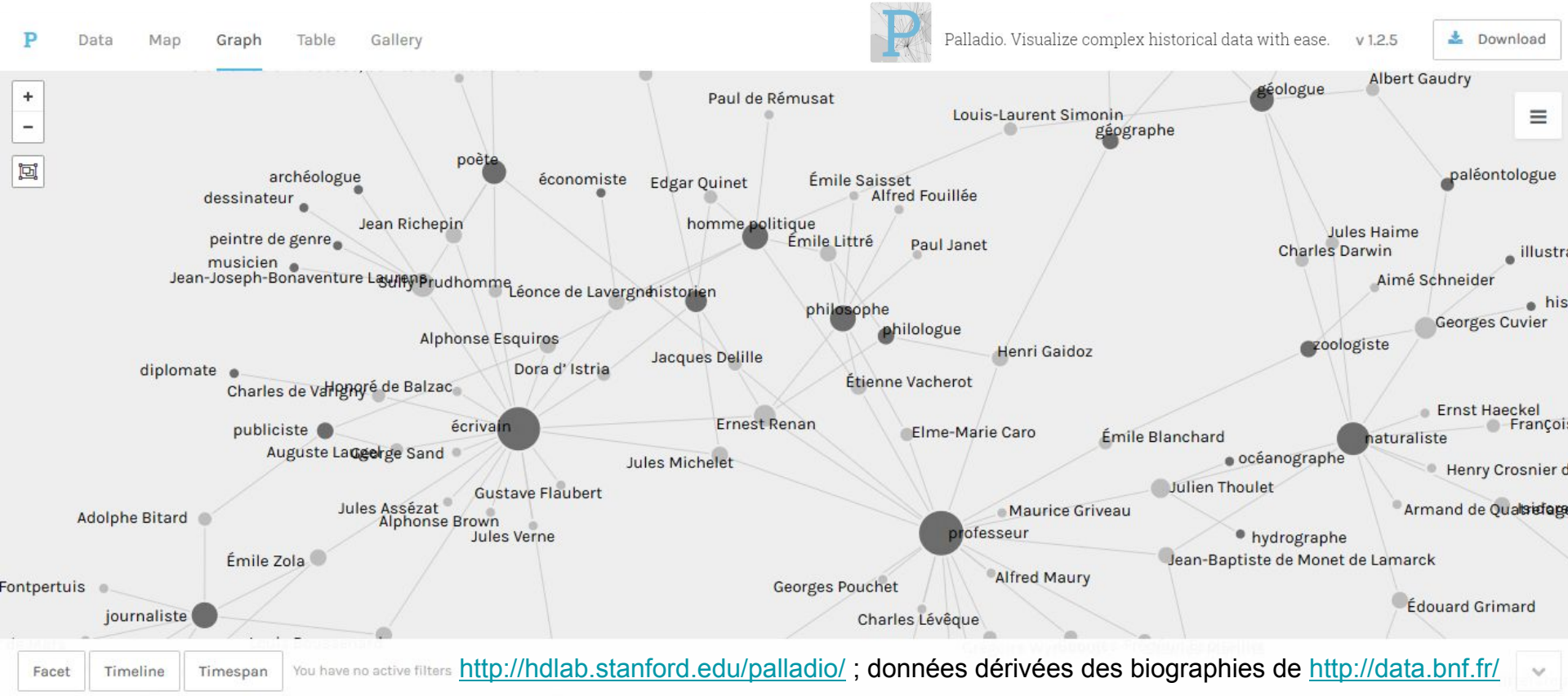
poétique des mystères de la vie

Cartographier le corpus

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

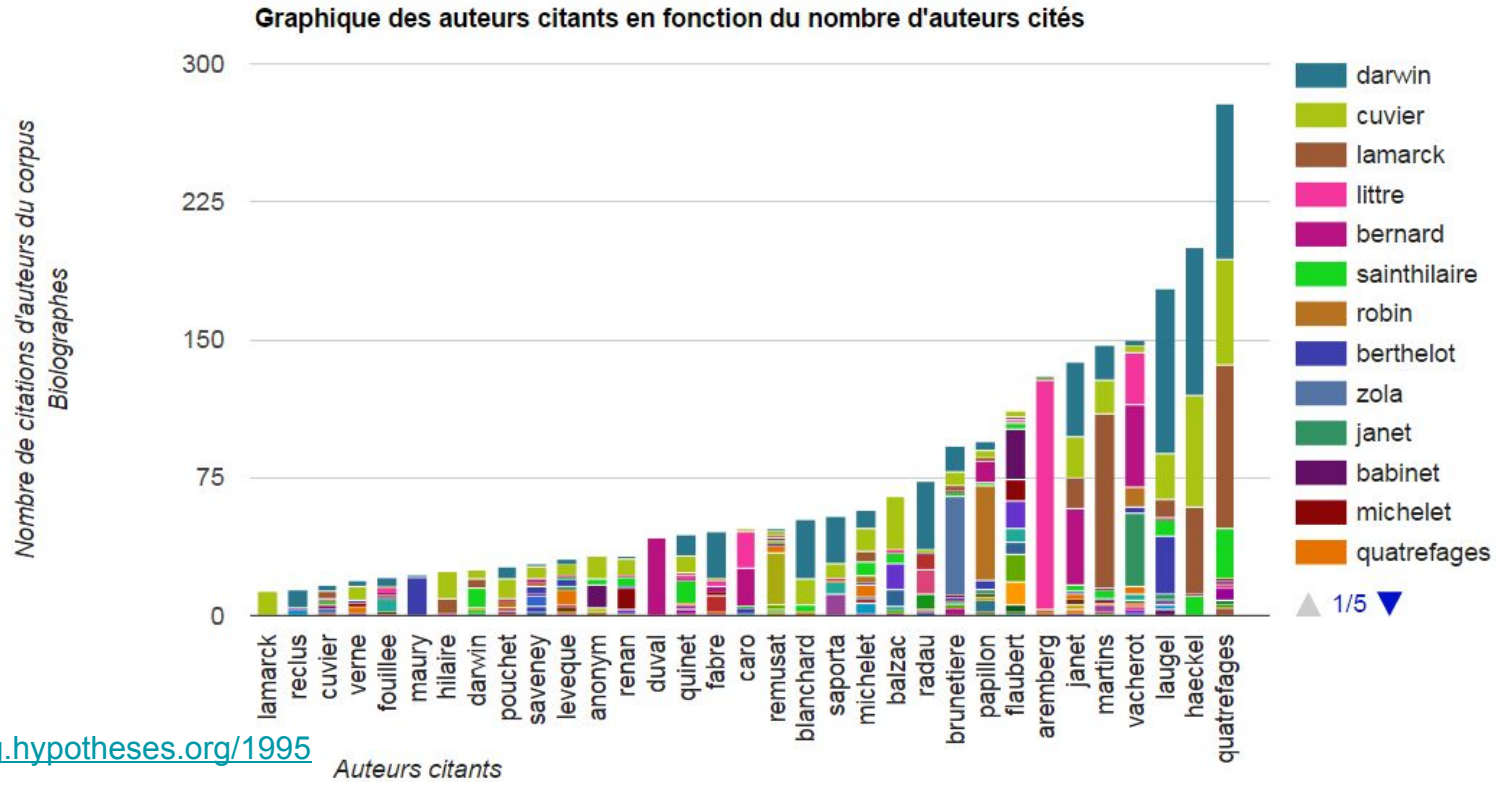
http://gallica.bnf.fr/ark:/12148/bpt6k1228258/f1

Données sur les auteurs



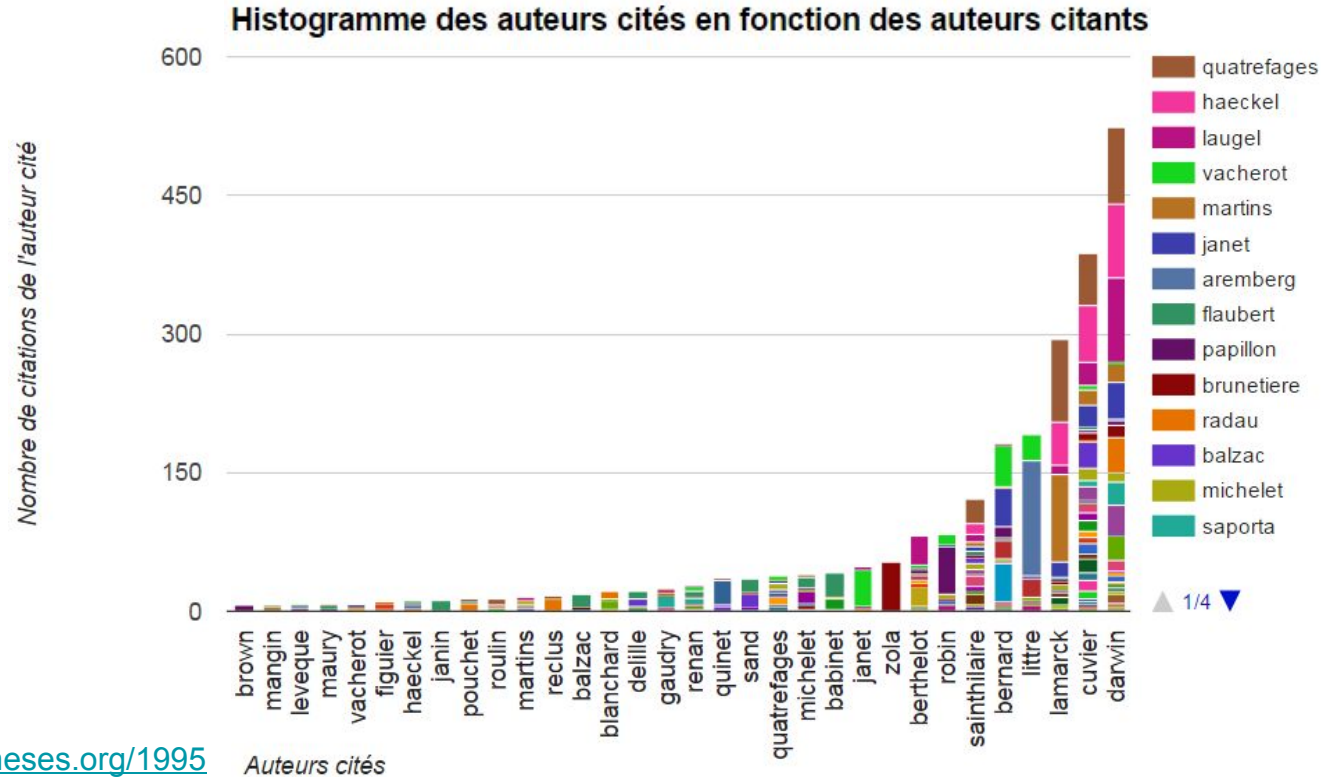
http://hdlab.stanford.edu/palladio/ ; données dérivées des biographies de http://data.bnf.fr/

Visualisation des citations entre auteurs du corpus



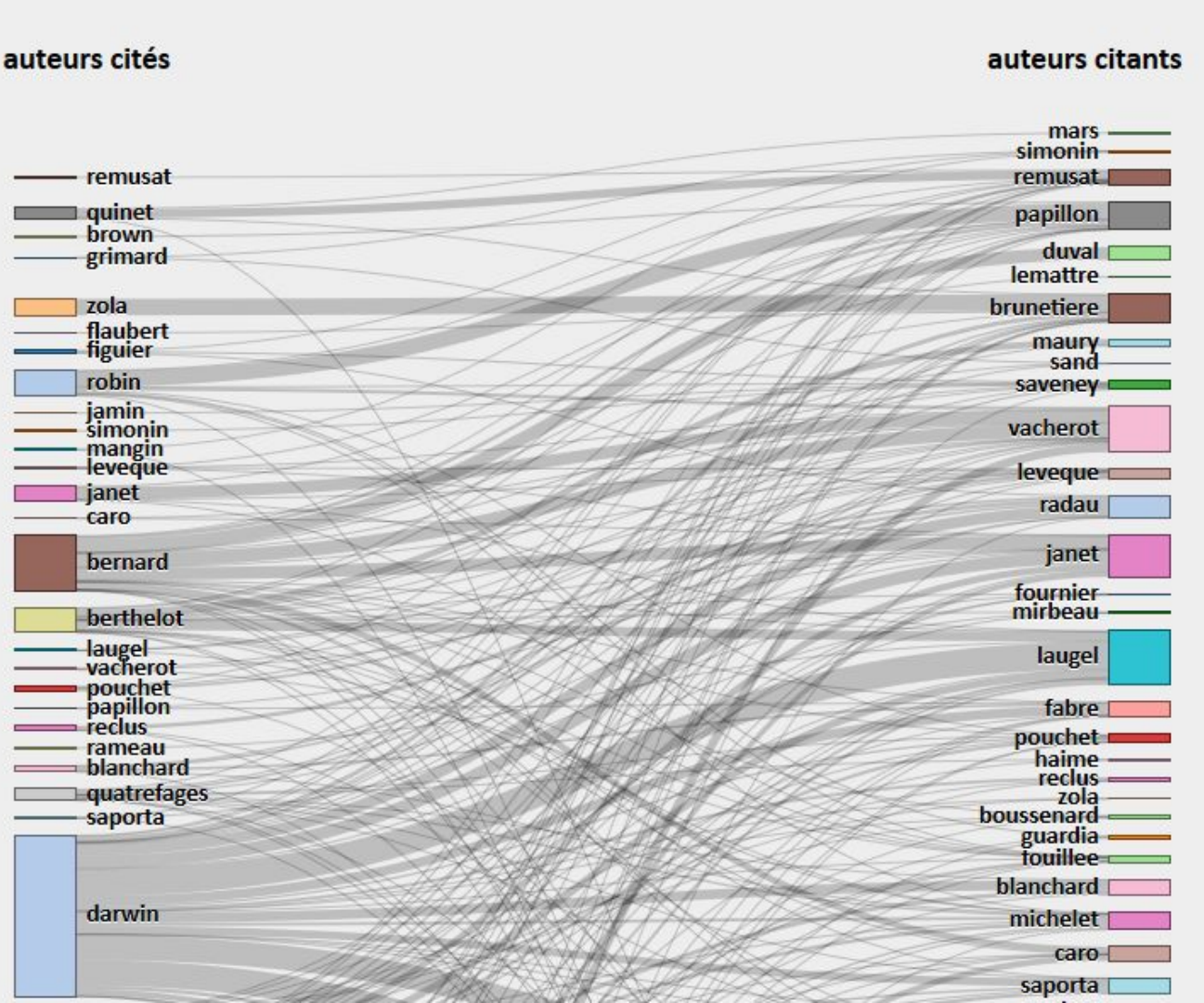
https://biolog.hypotheses.org/1995

Visualisation des citations entre auteurs du corpus



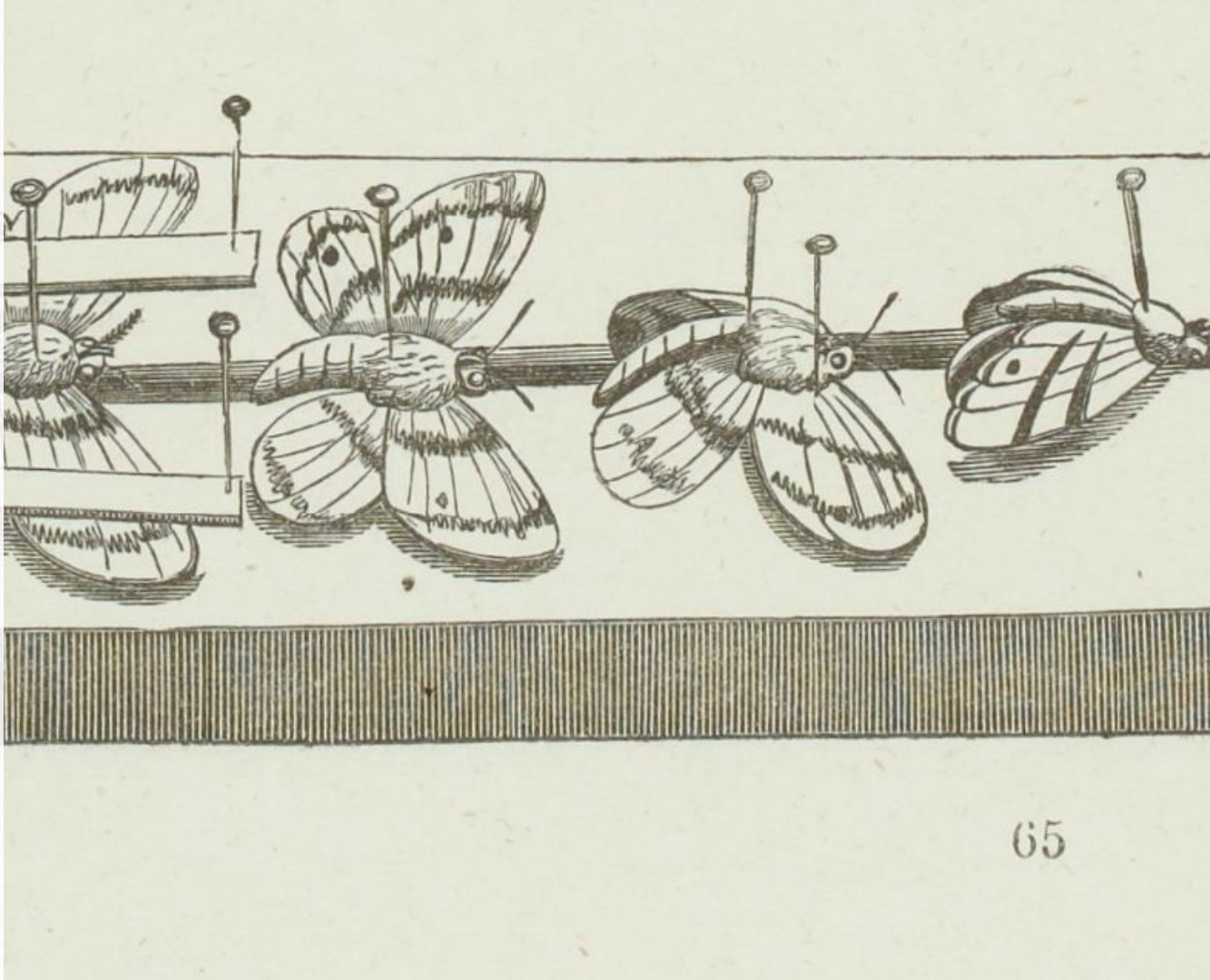
https://biolog.hypotheses.org/1995

Visualisation des citations



https://biolog.hypotheses.org/1995

Archiver le corpus



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

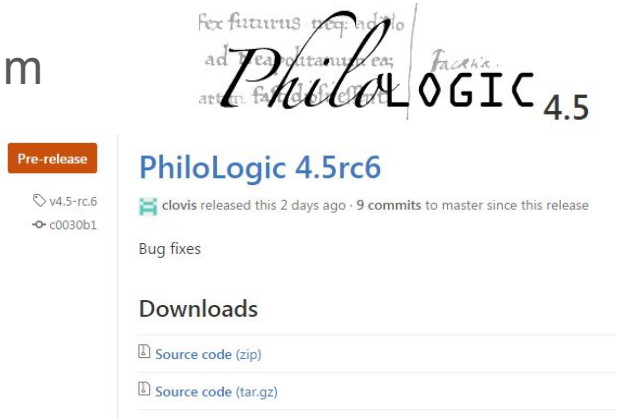
http://gallica.bnf.fr/ark:/12148/btv1b86002237/f279

Archivage avec HumaNum

Huma-Num : la très grande infrastructure des humanités numériques

Mise à disposition pérenne du corpus Biolographes 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



https://github.com/ARTFL-Project/PhiloLogic4
Démo sur http://artflsrv02.uchicago.edu/philologic4/encyclopedie/

Formatage des données pour Nakala

EXPOSER
Une plateforme pour déposer, documenter et diffuser vos données
nakala

Entrée :

1 fichier texte

Base de données du corpus Biolographes

Script en langage Python

Sortie pour Nakala :

1 fichier XML-TEI

1 fichier de métadonnées associé pour Nakala

Format Nakala

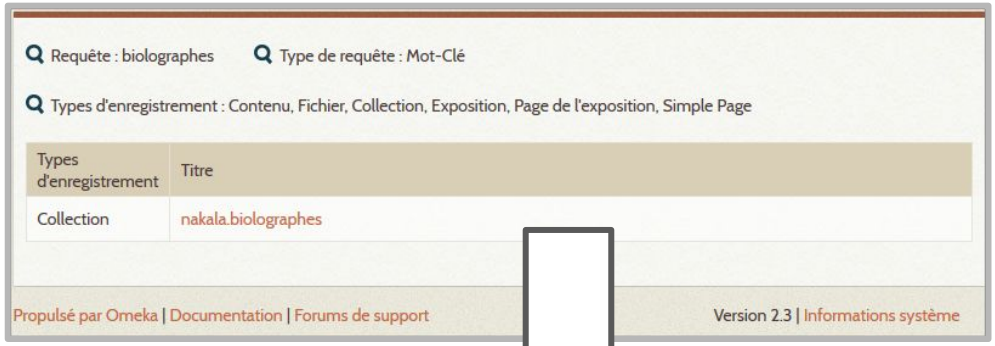
1 fichier XML-TEI

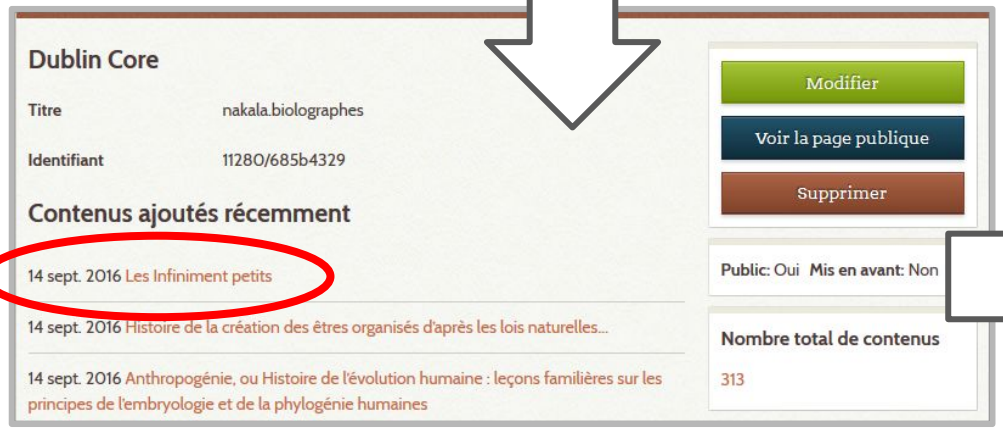
```
<TEI version="5.0"
xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title>Les Infiniment petits</title>
    <author>Anonyme</author>
    <date>1873</date>
   </titleStmt>
    <sourceDesc>https://fr.wikisource.org/
        wiki/Les_Infiniment_petits
    </sourceDesc>
  </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:identifier>https://fr.wikisource.org/wiki/Les_Infiniment_petits
   </dcterms:identifier>
   <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)





En conclusion

Un corpus :

- **évolutif** : version 2.0, ajouts de textes et d'auteurs
- **outillé** : logiciels TXM, TreeCloud, site http://corpus.biolographes.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 *Biolographes*
- 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 *Biolographes* :
  - 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 *Biolographes*
- Philippe Gambette, Nadège Lechevrel (2016), Explorer les réseaux de citations du corpus Biolographes, billet du carnet *Biolographes*