

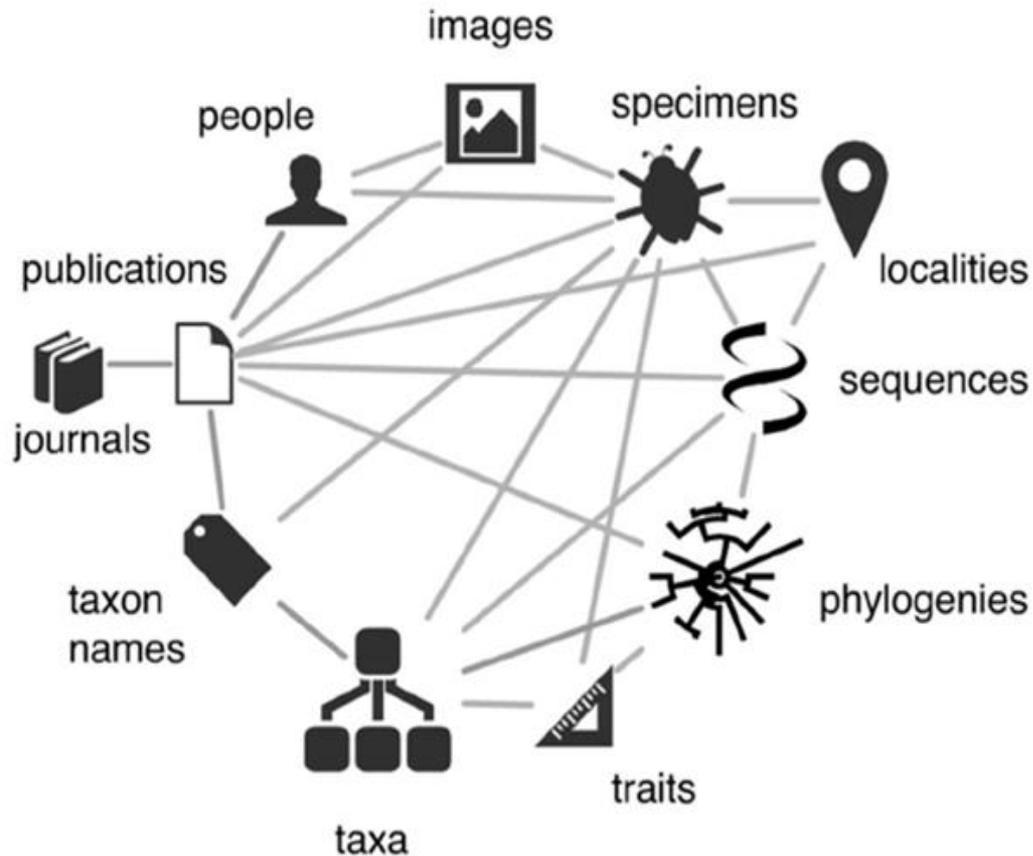
# Datapoc

Chloé Besombes, Simon Chagnoux

Journées professionnelles Collex-Persée  
Vendredi 5 avril 2019



# L'économie de la recherche en taxonomie à partir des collections naturalistes



# La cartographie

## Collections de spécimens

Herbier

Myriapodes

Mammifères

Ichtyologie

Géologie

Gouvernance: Direction des collections naturalistes (DGD-C)

## Bibliothèques

Sudoc-SiGB

Calames

HAL- Muséum

Gouvernance: Direction des bibliothèques (DGD-C)

## Publications scientifiques

Gouvernance: Service des Publications scientifiques (DGD-MJZ)

## Observations

Gouvernance: Inventaire national du patrimoine naturel (INPN)

## Listes nationales Référentiels

# La cartographie

---

- ▶ Types de données communes (noms scientifiques des espèces, noms de personnes, localisation géographique, dates)
- ▶ Pratiques de production et de normalisation hétérogènes
- ▶ Des silos étanches entre eux mais ouverts sur des axes ou internationaux



# Les référentiels au MNHN

---

- ▶ Publié par le MNHN: Taxref

- ▶ <https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref>

- ▶ Alimentés (exemples):

- ▶ <http://www.marinespecies.org/>

- ▶ <https://www.worldbirdnames.org/>

- ▶ <https://www.fishbase.in/search.php>

- ▶ <http://www.ipni.org/>

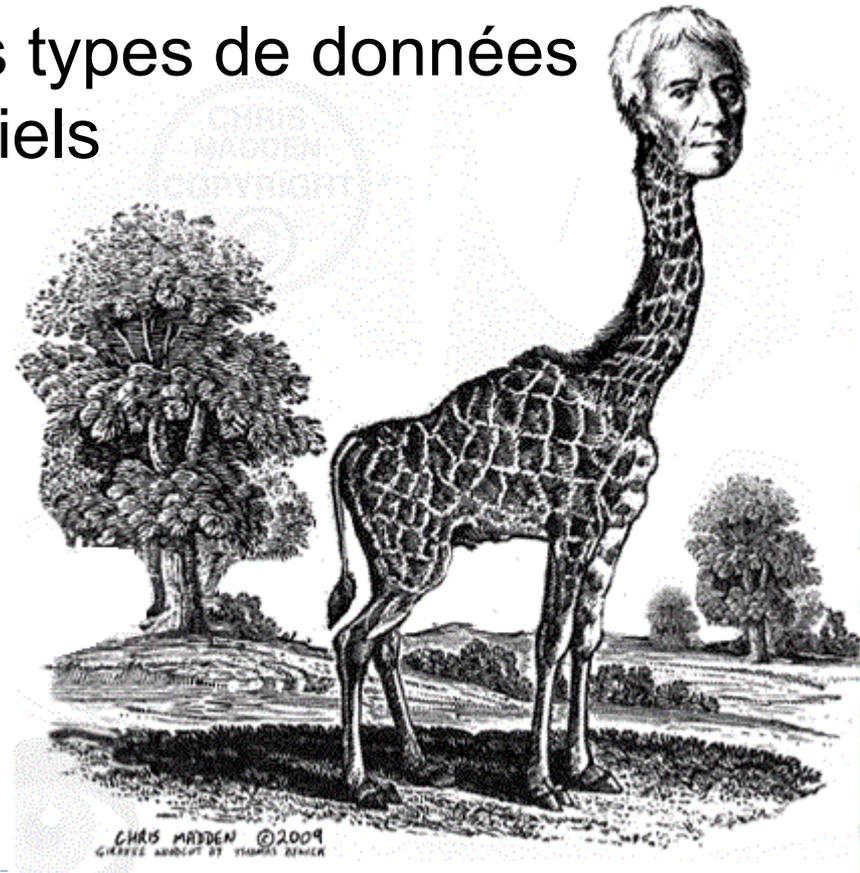
- ▶ <https://www.idref.fr/>

- ▶ .....



# Le pivot: les personnes

- ▶ Présentes dans tous les systèmes d'informations
- ▶ Stables par rapport à d'autres types de données structurées dans des référentiels
- ▶ Hétérogènes selon les bases



# Les personnes dans les bases MNHN

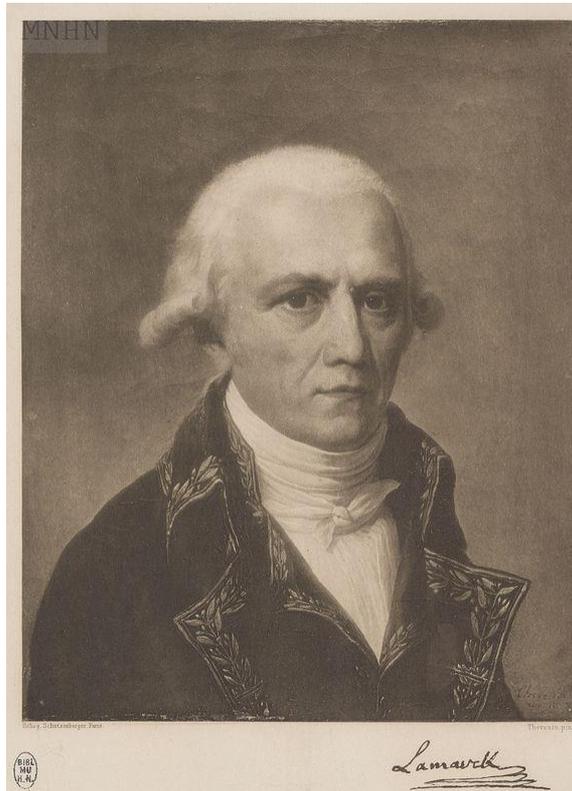
Base de données de  
spécimens  
Pas d'identifiants

Bibliothèques

ISNI, ORCID,  
Idref, IdHAL...

Publications  
scientifiques

scopusid,  
ResearchID, VIAF ou  
ORCID...



[PO2034](#) – Bibliothèques du  
MNHN

INPN

Pas d'identifiants



# Le corpus

- ▶ 467 noms
- ▶ Taxonomistes
- ▶ XVII<sup>e</sup> – XXI<sup>e</sup> s



Source gallica.bnf.fr / Bibliothèque nationale de France

*Du jardin au Muséum : en 516 biographies*, Philippe Jaussaud & Edouard-Raoul, Publications du Muséum national d'Histoire naturelle, 2004



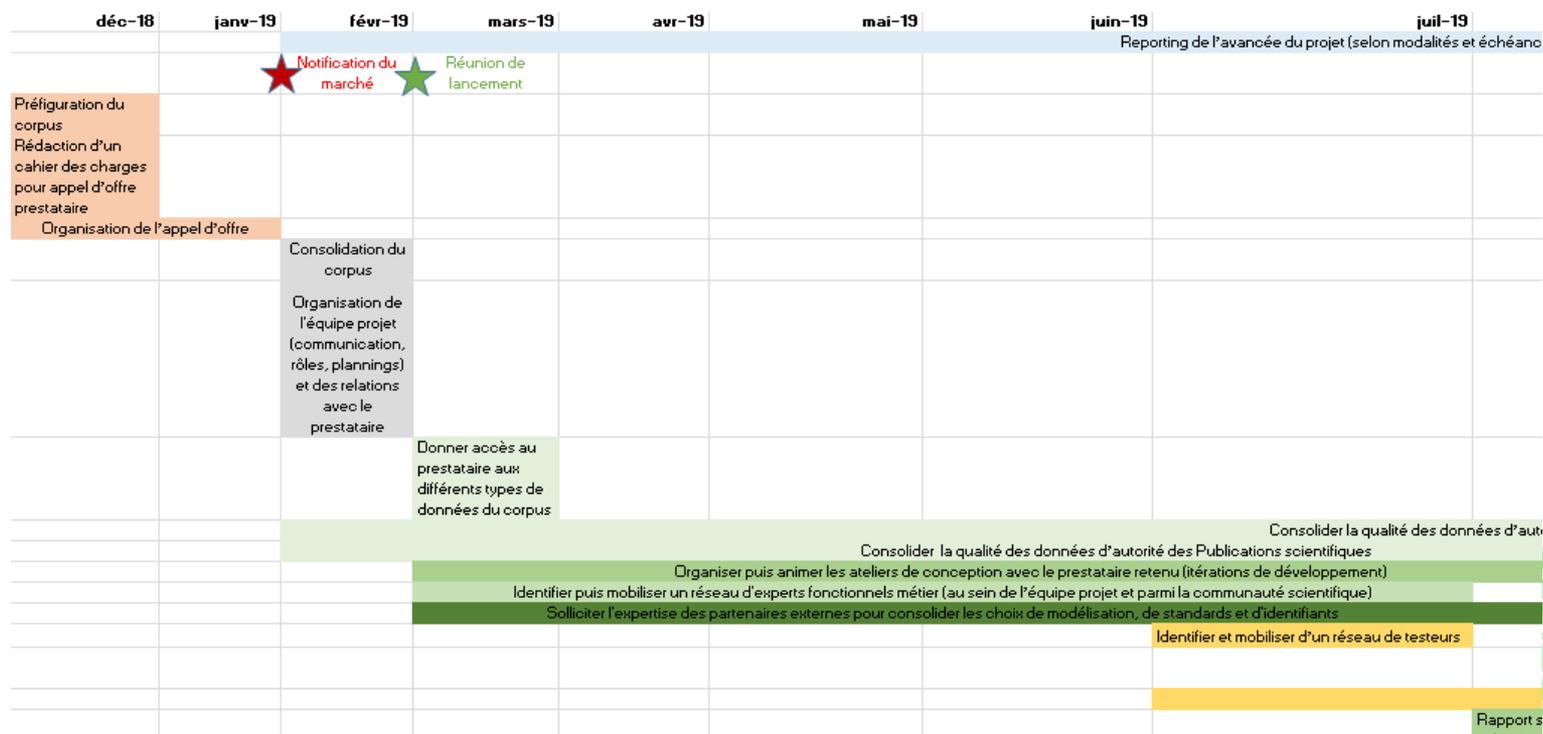
# Représentation du corpus dans Idref

---

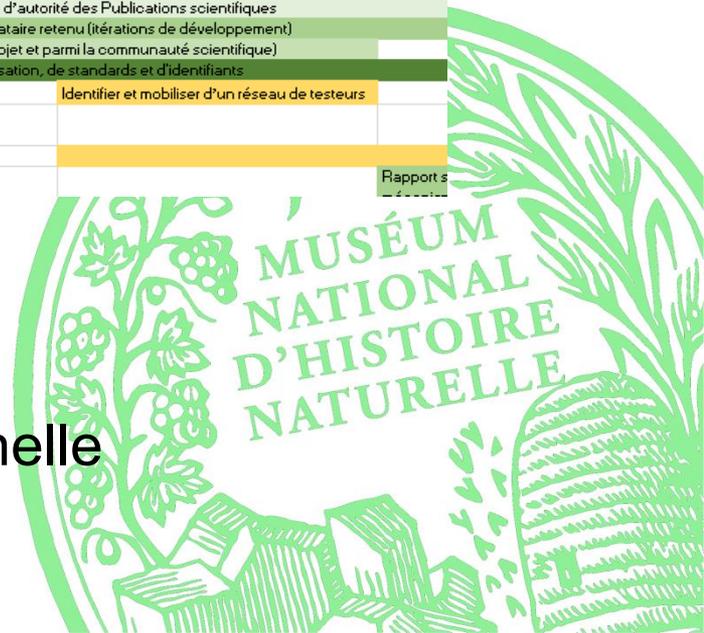
- ▶ Premier travail d'alignement en collaboration avec l'Abes
- ▶ 85% de taux de couverture dans IdRef
- ▶ Ouverture d'un chantier qualité des données sur le corpus
  - ▶ Autorités
  - ▶ Liens aux notices bibliographiques
  - ▶ Recherche des 15% restants



# Le projet



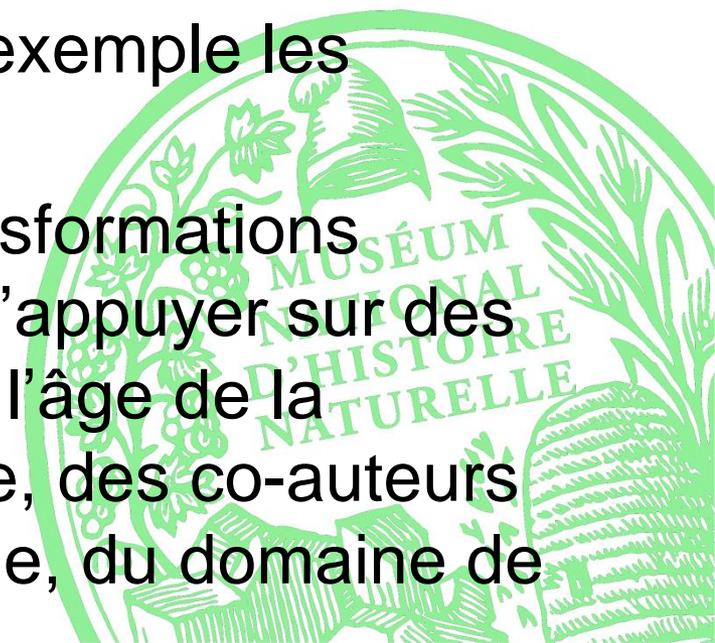
- ▶ L1 : Jeu de données structurées
- ▶ L2 : Site et API
- ▶ L3 : Rapport sur les mécanismes
- ▶ L4 : Préconisations pour passage à l'échelle



# Les mécanismes

---

- ▶ 100 000 « récolteurs » différents pour 9 millions d'objets
- ▶ Unifier les noms de personne même lorsque les présentations sont différentes (« Jean-Baptiste Lamarck », « Lam. », « Lamarck, chevalier de Monet », « Lamarck, JB », « Lamarck », ...)
- ▶ Désambiguïser les homonymes (exemple les « Jussieu »)
- ▶ Ces méthodes utiliseront des transformations syntaxiques mais peuvent aussi s'appuyer sur des heuristiques : prise en compte de l'âge de la personne au moment de la récolte, des co-auteurs d'une publication, de la géographie, du domaine de compétence, etc.



# Le site « people of collection »

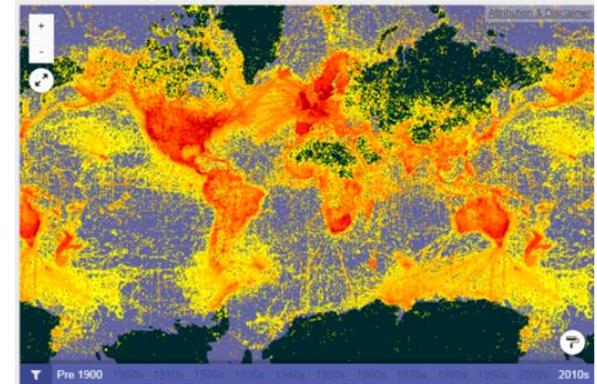
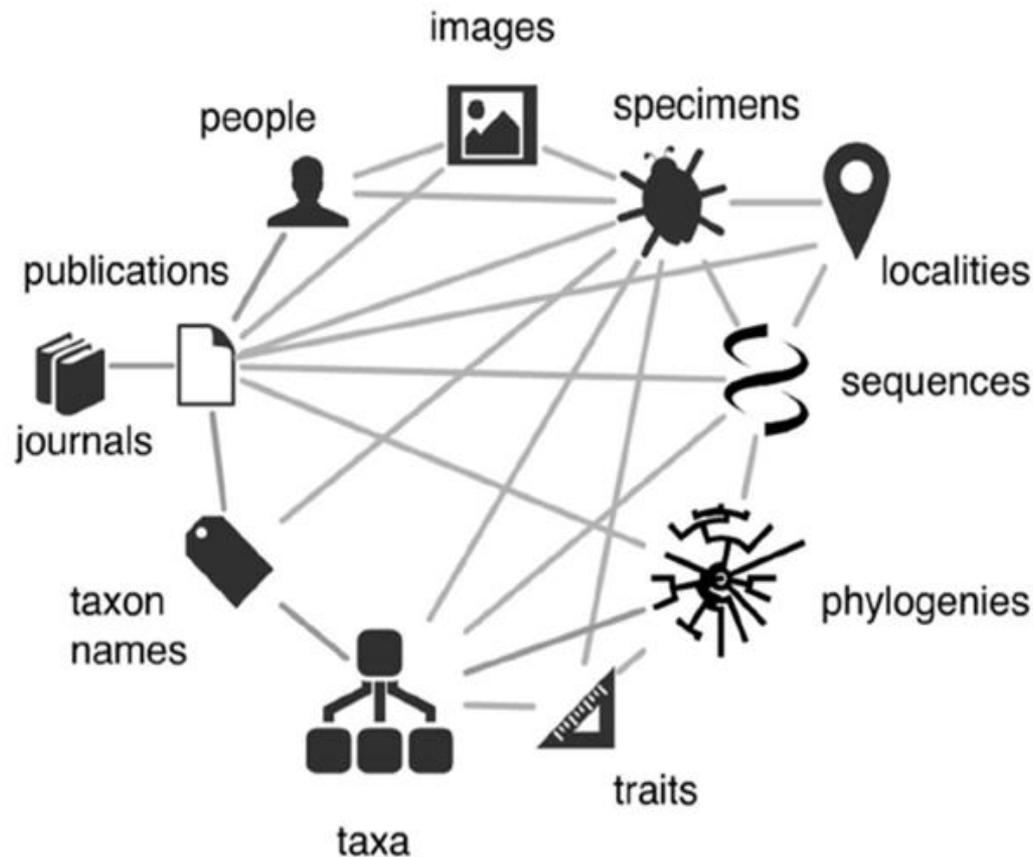
The screenshot shows the website interface for the 'people of collection' project. At the top, there is a header for 'MUSÉUM NATIONAL D'HISTOIRE NATURELLE' with a search bar and navigation links. Below the header is a large image of two swans. The main content area is titled 'JEAN-BAPTISTE DE LAMARCK' and displays a list of his works. Each work entry includes a small image of the book cover, the title, the date of publication, the publisher, and the location of the manuscript. The works listed include 'SYSTÈME ANALYTIQUE DES CONNAISSANCES POSSESSES DE ENCHÔME, RESTRIENTES À CELLES QUI PROVIENNENT DIRECTEMENT OU INDIRECTEMENT DE L'OBSERVATION', 'DE L'INFLUENCE DE LA LUNE SUR L'ATMOSPHÈRE TERRESTRE', 'PLANCHES ORIGINALES DE LA BOTANIQUE', 'ENCYCLOPÈDE MÉTHODIQUE. BOTANIQUE. VOL 2', 'PHILOSOPHIE GÉOLOGIQUE, OU EXPOSITION DES CONSIDÉRATIONS RELATIVES À L'HISTOIRE NATURELLE DES ANIMAUX', and 'BOTANICA'. On the left side, there are navigation menus for 'COLLECTIONS', 'ŒUVRES', 'ARTICLES', 'IMAGES', 'BIBLIOGRAPHIE', 'LIEUX DE CONSERVATION', 'ANNÉE DE PUBLICATION', 'COLLECTION', 'THÉMATIQUES', and 'RECHERCHES ASSOCIÉES'. At the bottom, there is a section for 'RECHERCHES ASSOCIÉES' featuring portraits of other naturalists: Jean-Jacques Auvouart (1764 - 1834), Leclerc de Buffon (1732 - 1788), Charles-Alexandre Lesueur (1758 - 1844), and Georges Cuvier (1769 - 1832).

- ▶ Une page pour chaque Naturaliste



# L'accès ouvert au données

## ▶ API ou ressources RDF



# L'évaluation par groupe utilisateurs



# Le passage à l'échelle

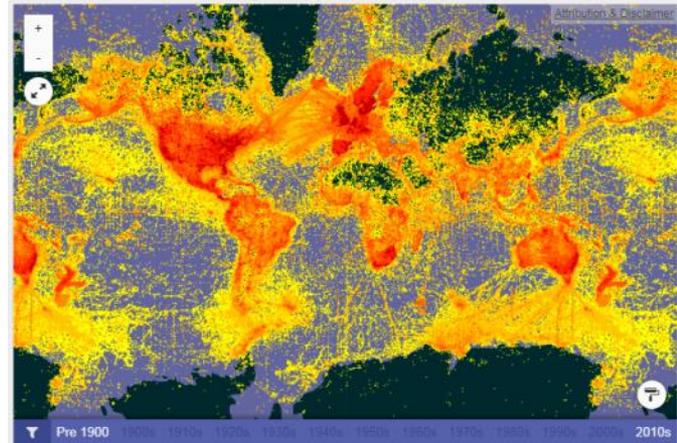
---

- ▶ Mécanismes scriptés et rejouables
- ▶ Estimation de la fiabilité des heuristiques
- ▶ Evaluation du cout humain de validation
- ▶ Mesure de l'impact sur les différents silos



# Faire évoluer le SI Global de la biodiversité

- ▶ 1.1 milliard d'occurrences
- ▶ 43 421 jeux de données
- ▶ 1378 institutions
- ▶ Darwin Core



recordedBy  
georeferencedBy  
identifiedBy  
scientificNameAuthorship  
nameAccordingTo



# Merci pour votre attention

---

- ▶ [chloe.besombes at mnhn.fr](mailto:chloe.besombes@mnhn.fr)
- ▶ [simon.chagnoux at mnhn.fr](mailto:simon.chagnoux@mnhn.fr)

