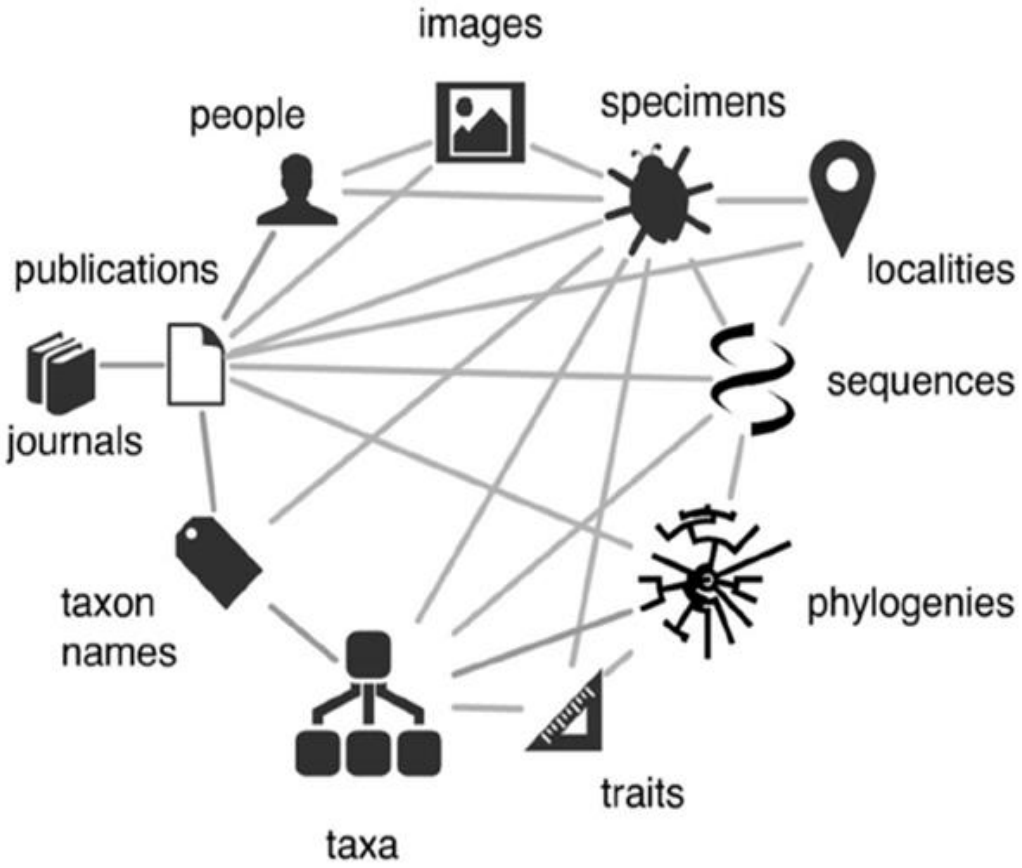
Datapoc

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

Listes nationales Référentiels

Gouvernance: Inventaire national du patrimoine naturel (INPN)

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 [illegible]ux ou internationaux

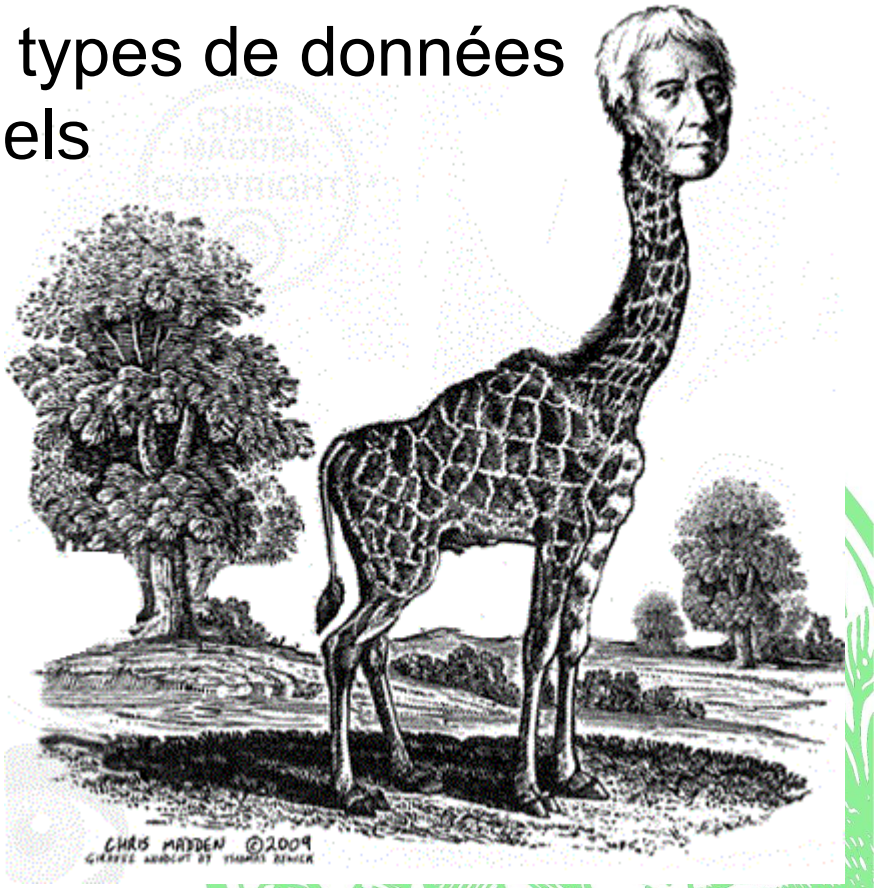
Les référentiels au MNHN

- 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

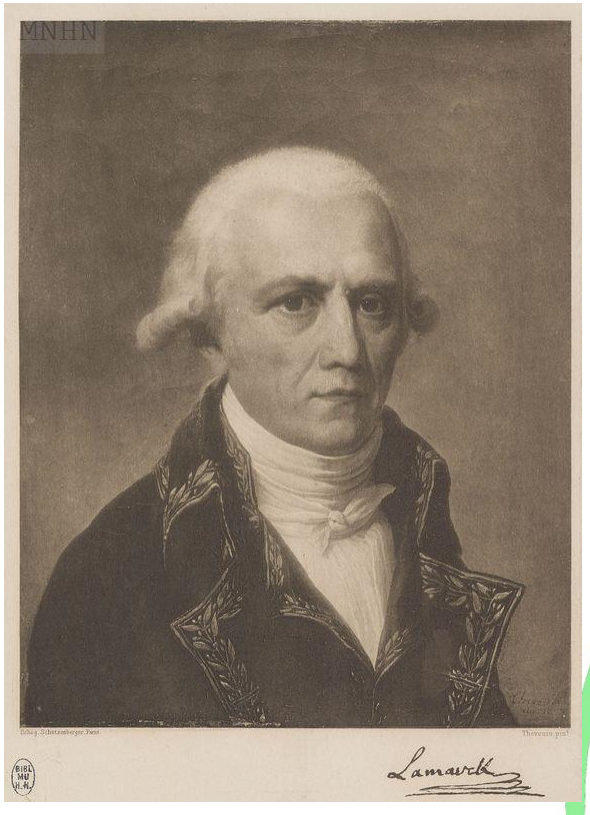
Base de données de spécimens
Pas d'identifiants

Bibliothèques

ISNI, ORCID, Idref, IdHAL…

Publications scientifiques

ScopusID, ResearcherID, VIAF ou ORCID…



PO2034 – Bibliothèques du MNHN

INPN

Pas d'identifiants

Le corpus

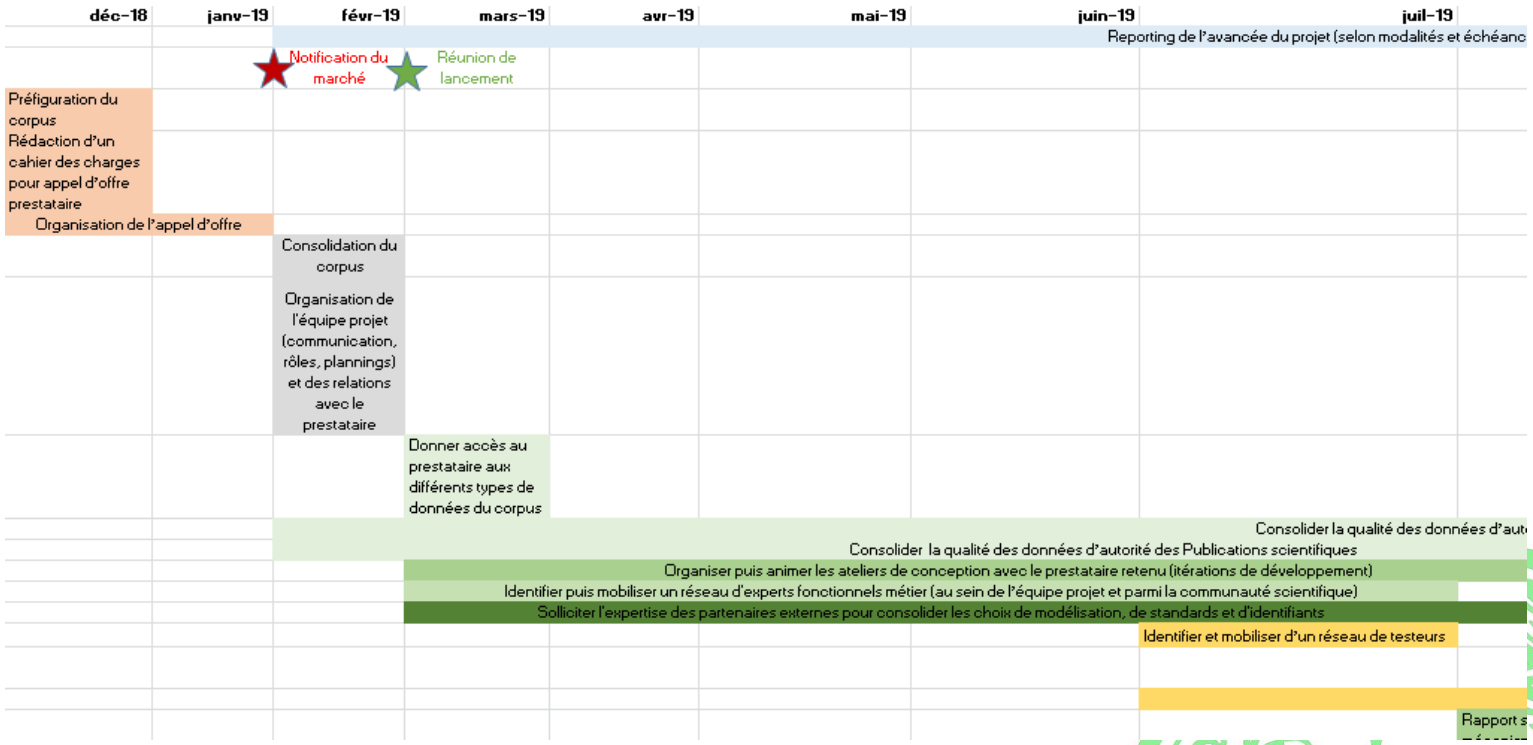
- 467 noms
- Taxonomistes
- XVII^{e} – XXI^{e} s

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

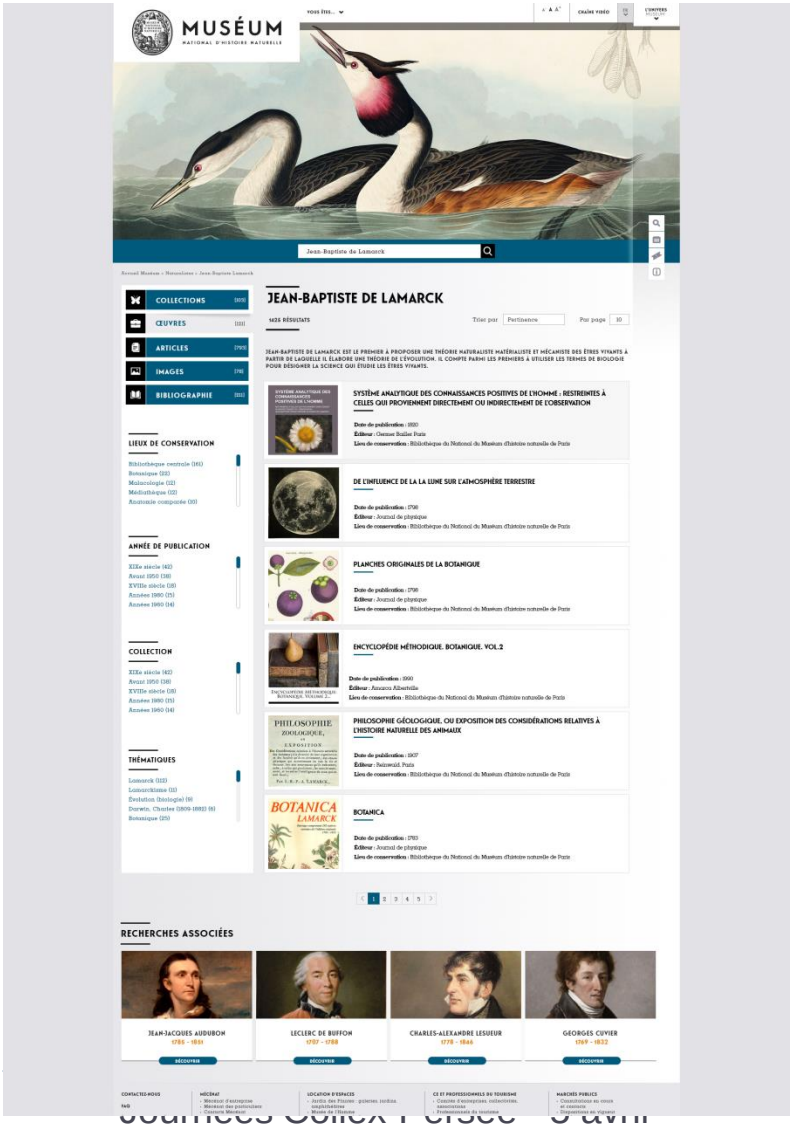


- L1 : Jeu de données structurées
- L2 : Site et API
- L3 : Rapport sur les mécanismes
- L4 : Préconisations pour passage à l'échelle

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.

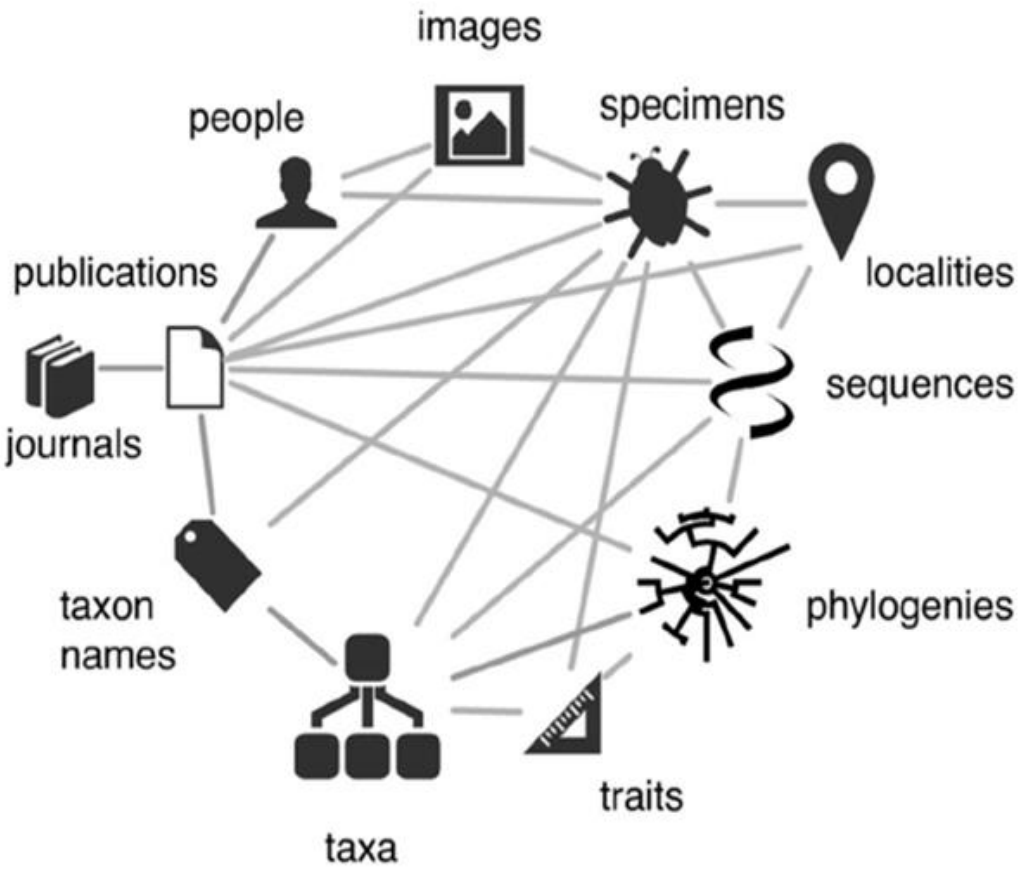
Le site « people of collection »

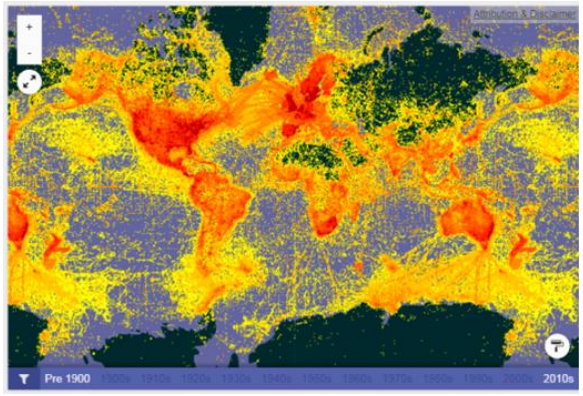


- Une page pour chaque Naturaliste

L’accès ouvert au données

- API ou ressources RDF





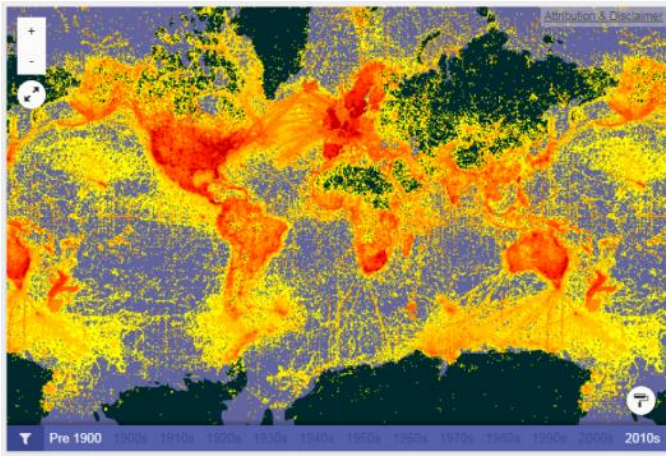
L’évaluation par groupe utilisateurs

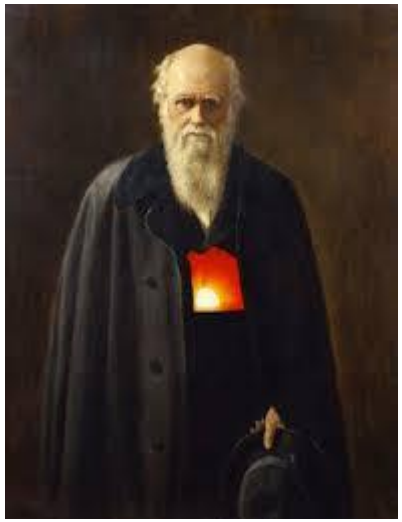
Le passage à l'échelle

- 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

MUSÉUM NATIONAL D'HISTOIRE NATURELLE

Merci pour votre attention

- chloe.besombes at mnhn.fr
- simon.chagnoux at mnhn.fr