

Harmonisation des noms de taxons pour les données de biodiversité

généralités et zoom sur les plantes

Matthias Grenié, Emilio Berti, Juan Carvajal-Quintero,
Alban Sagouis, Gala Mona Louise Dädlow,
et Marten Winter

24 Mai 2022 – Atelier D2KB



Présentation basée sur un article publié

Methods in Ecology and Evolution



REVIEW | Open Access |

Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices

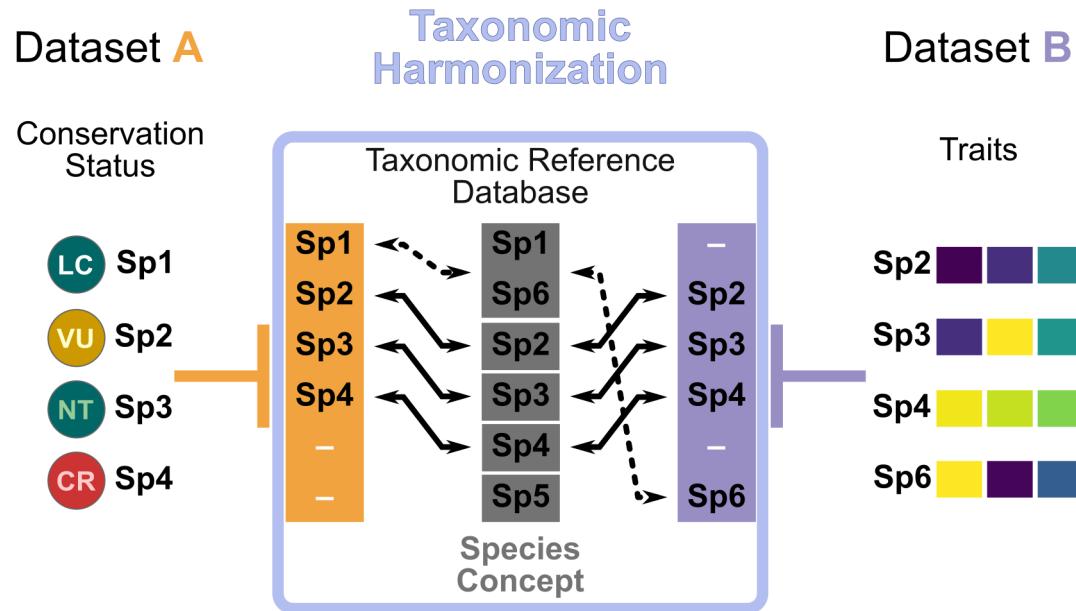
Matthias Grenié , Emilio Berti, Juan Carvajal-Quintero, Gala Mona Louise Dädlow, Alban Sagouis, Marten Winter

First published: 17 January 2022 | <https://doi.org/10.1111/2041-210X.13802>

<https://link.infini.fr/harmo>

Vous avez parlé d' « harmonisation taxinomique » ?

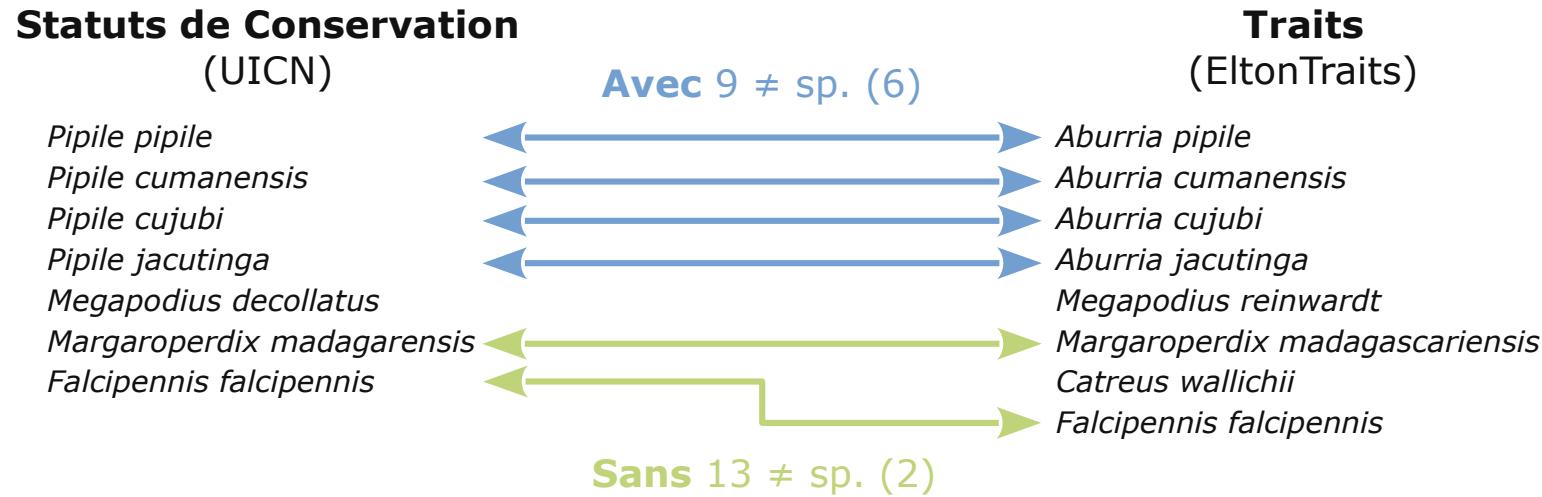
Macroécologie = correspondance de jeux de données pour répondre à des questions



L'harmonisation taxinomique **homogénéise** des noms de taxon grâce à une **liste de référence**

Pourquoi harmoniser des noms de taxons ?

Exemple : Risque d'extinction d'oiseaux en fonction de leur biomasse



L'**harmonisation taxinomique** permet la **comparaison**

L'harmonisation taxinomique : un paysage brumeux



- Pas de perspectives générales sur les **sources**
- Pas de vue d'ensemble des **outils disponibles**
- Pas de **guide** sur comment procéder

Les (Macro-)écologues ont besoin d'un **guide** !

→ notre motivation: écrire un **article de synthèse**



Décrire le paysage

Sources de taxinomie : les bases de données

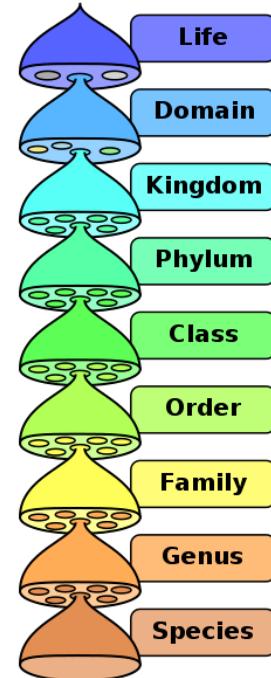
Taxonomic Reference Database
(= Base de données de Référence Taxinomique)

(aussi taxonomic backbone,
taxonomic checklist, taxonomic authority, etc.)

« Base de données centralisées d'information nomenclaturelle »

+ Correction d'erreur

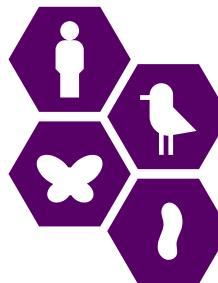
+ Résolution
synonymique



Sources de taxinomie : exemples de bases de données



Backbone
GBIF (2021)



Sans restriction
taxinomique

Juin 2021
6.6M de noms
3.7M acceptés
2.6M synonymes

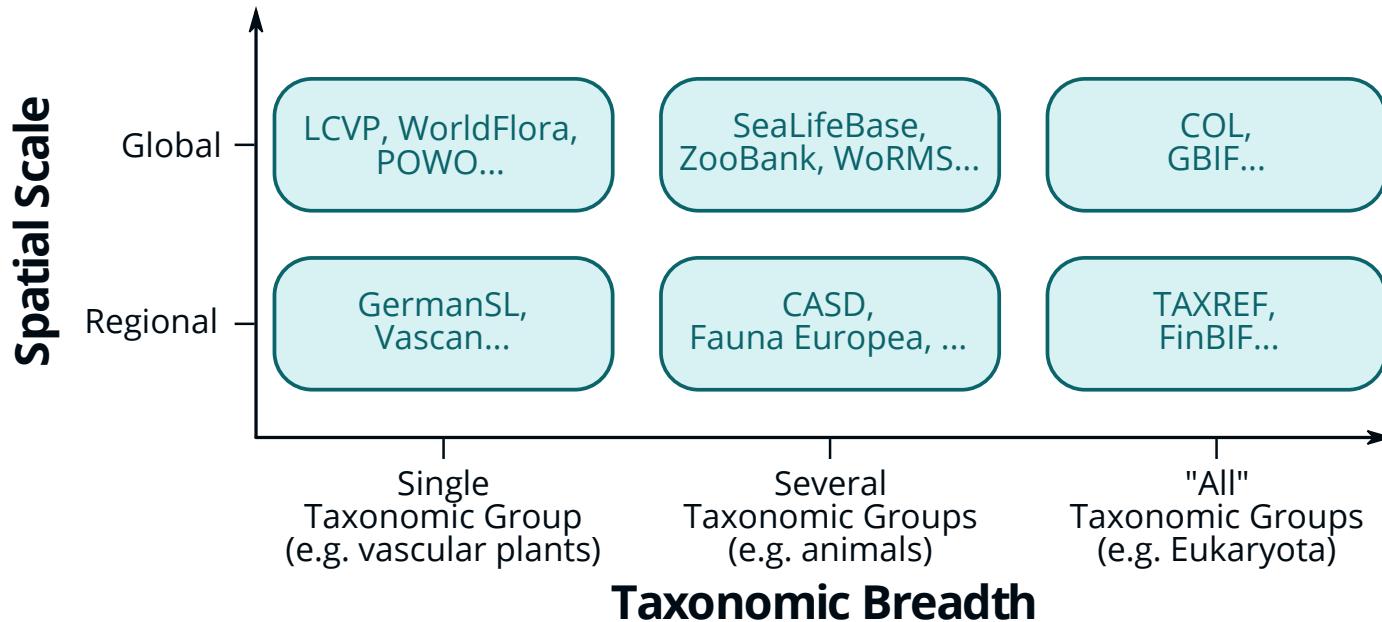
Leipzig Catalogue
of Vascular Plants
LCVP
(Freiberg et al. 2020)



Plantes
Vasculaires

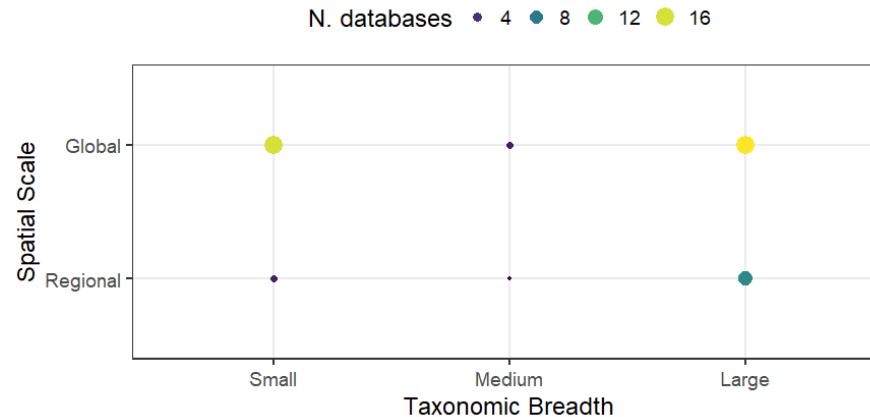
Nov. 2020
1.3M de noms
350k acceptés
850k synonymes

Sources de taxinomie : typologie des bases de données



Sources de taxinomie : liste de base de données

Identification de **50 bases de données**
(non exhaustif)



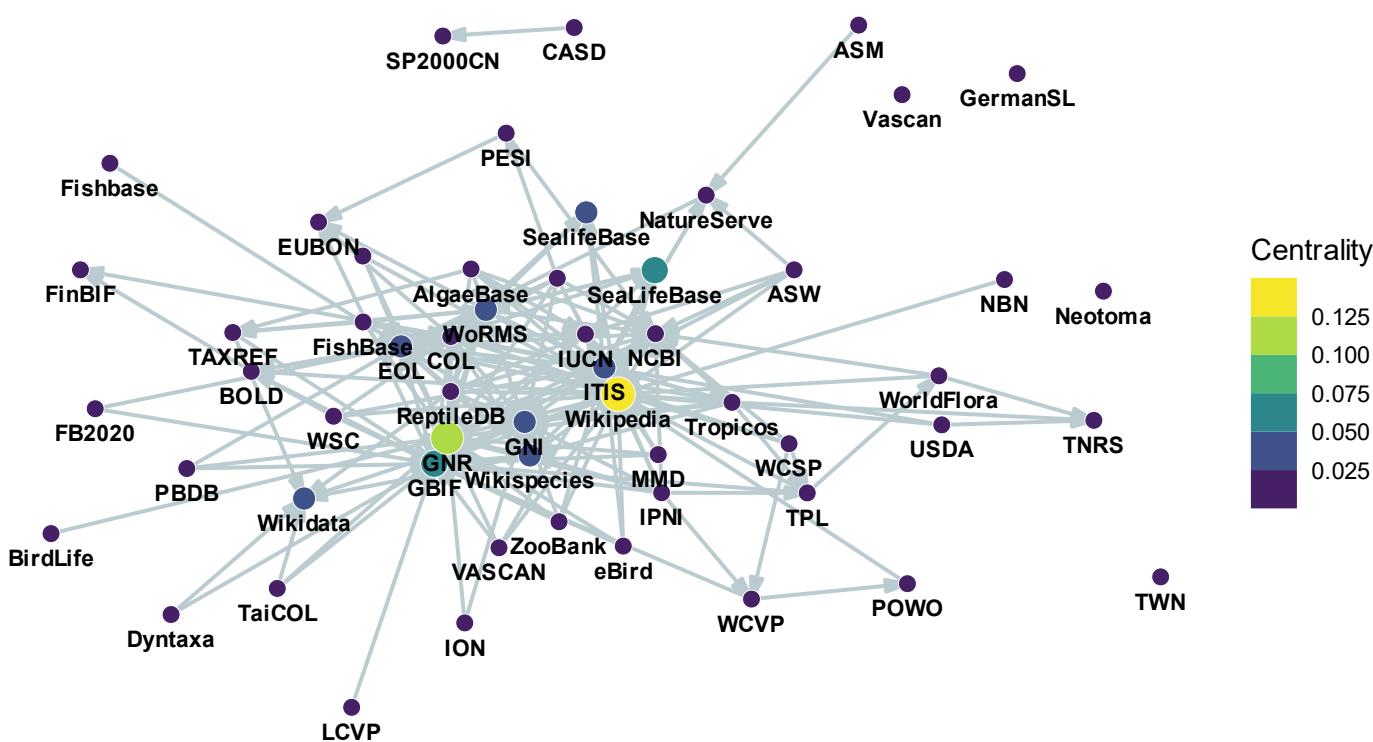
NOMBREUSES **Mondiales** de **petite** ou **large** étendue taxinomique

Sources de taxinomie : relation entre bases de données

Arête
Source → Cible

Quelques bases très réutilisées

Quelques bases « mega- agrégatrices »



N.B. : Cette information est **très difficile à obtenir**
(sans parler de la dimension quantitative)

A close-up photograph of a three-toed sloth hanging from a tree branch. The sloth's dark brown, almost black, fur is visible, along with its long, thick tail. It is holding a large, bright orange, textured flower cluster with its front paws and is eating from it. The background consists of various green leaves and branches, with sunlight filtering through the canopy.

Des bases de données aux outils

Outils taxinomiques : paquets R

Revue des paquets sur le CRAN + GitHub + Bioconductor



Tri manuel à partir de requêtes standard (recherche GitHub + r-pkg.org)

Critères

- **Vrai paquet R** (pas de liste de scripts)
- **Pas un « wrapper »** (outils originaux)
- **Ne traitant pas de génomique** (champ différent)

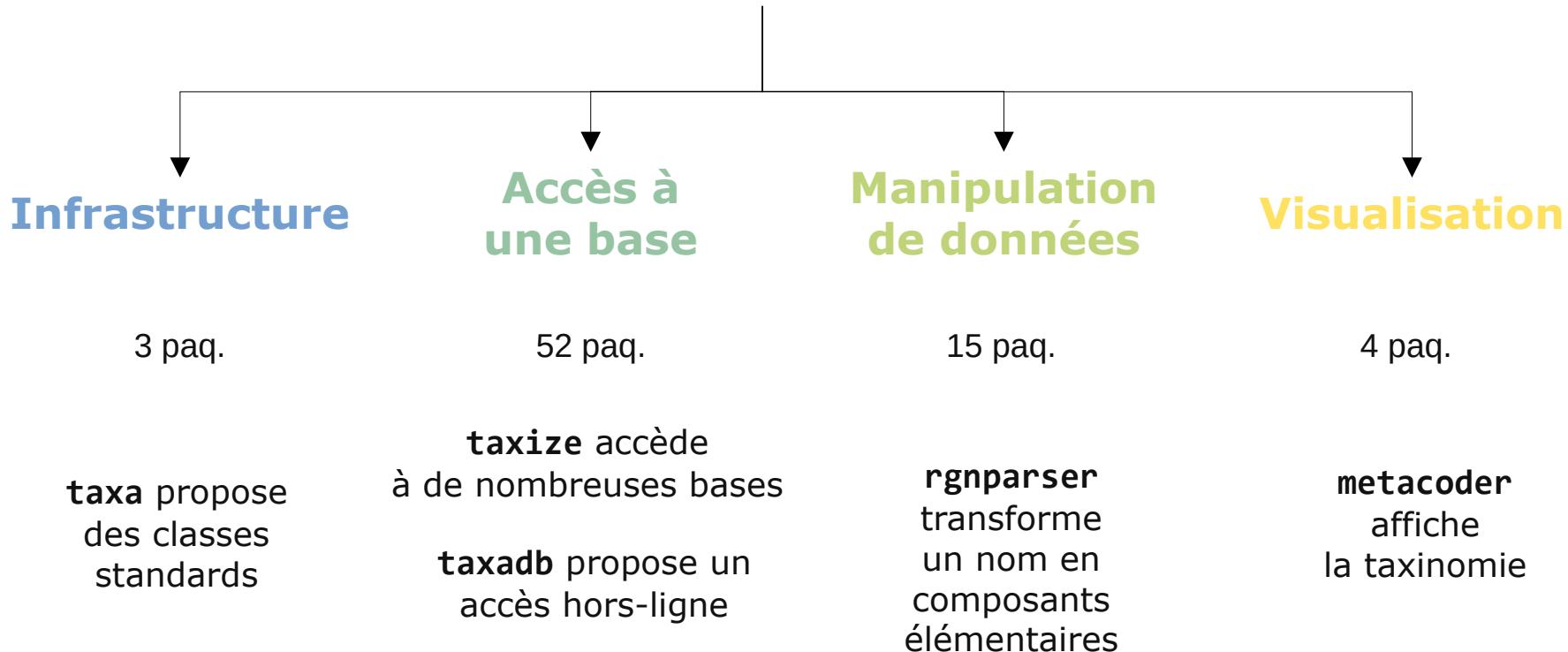
Identification de **73 paquets** dont **64** inclus

taxize



Outils taxinomiques : différentes catégories

64 paquets appartenant à **4 catégories**

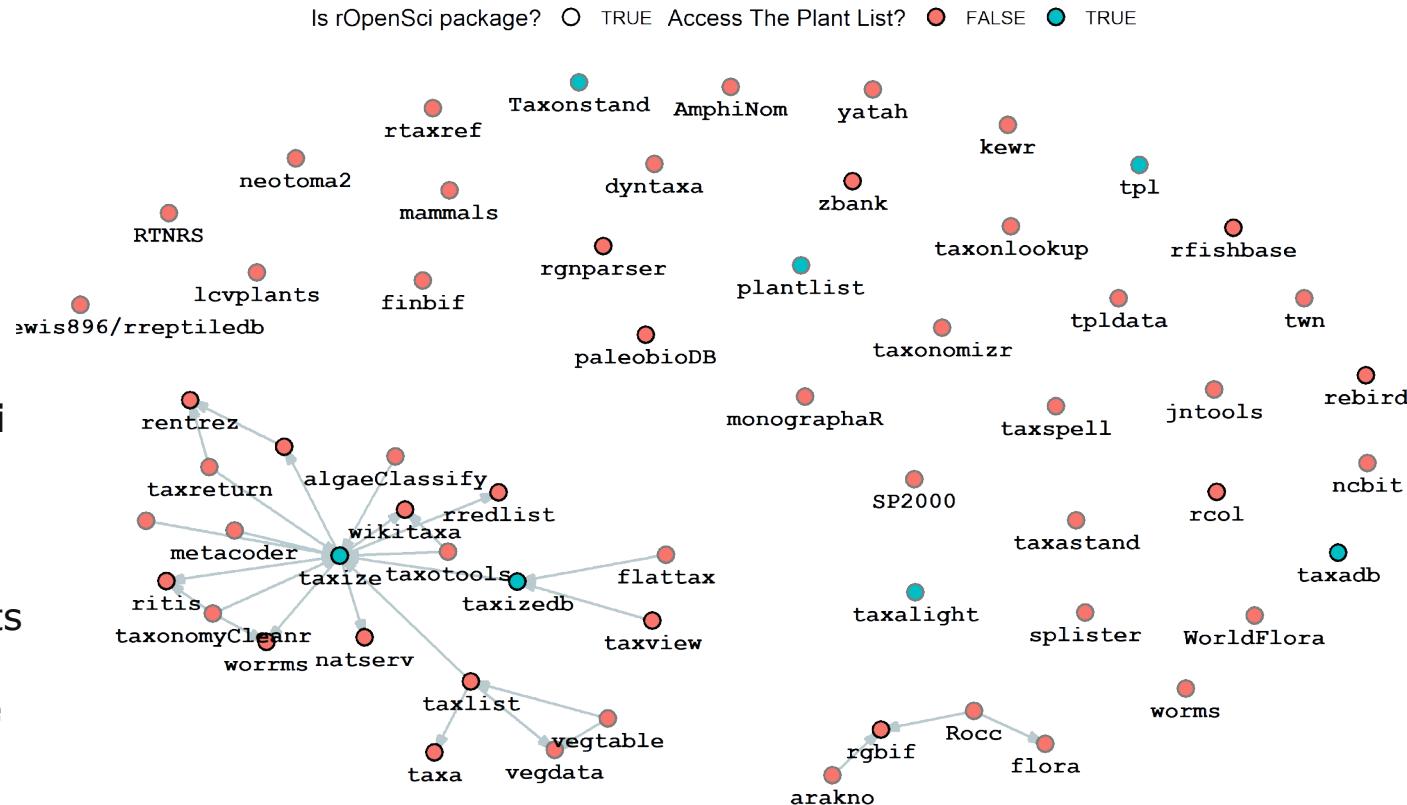


Réseau de dépendances entre paquets

Réseau très éclaté

Sauf les **paquets rOpenSci**

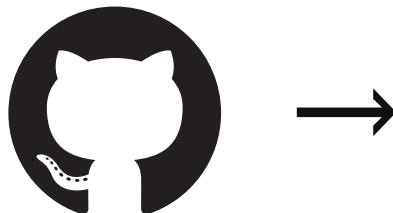
Beaucoup de paquets
accède
à la **même source**
(TPL)



Pas de standard pour manipuler les données taxinomiques

Outils taxinomiques : enseignements tirés

GitHub n'est **pas**
une **archive pérenne**



...

Contribuer
plutôt que **recréer**

Le dév. d'outils taxinomique est **DIFFICILE**
→ **Expertise rare** (& peu de motivation)



Cartographier le paysage des outils taxinomiques



taxharmonizexplorer

- Affiche les **relations dirigées** entre bases & paquets
- Affiche **information résumées** sur les nœuds
- Base évolutive
(mise-à-jour continuellement)

Accès via :
<https://tiny.cc/taxharmonizexplorer>



taxharmonizexplorer Description Network Help

Selected Node Information

Node Name: WorldFlora_pkg
Type: package
Actively Maintained: yes
Workflow Step(s):
Database Access
Release URL:
<https://cran.r-project.org/package=WorldFlora>

Click on one (several) node(s) to highlight it (them) in the network:

Show 10 entries Search:

Name	Type	Tax. Group
1 algaeclassify	package	phytoplankton
2 finbif	package	No taxonomic restriction
3 insect	package	No taxonomic restriction
4 lcvplants	package	land plants
5 metacoder	package	microbes
6 microclass	package	prokaryotes
7 monographaR	package	land plants
8 natServ	package	No taxonomic restriction
9 plantlist	package	plants
10 rcol	package	No taxonomic restriction

Showing 1 to 10 of 85 entries

Previous 2 3 4 5 ... 9 Next

Relationships between taxonomic R packages and databases

Legend:

- package (purple circle)
- database (orange triangle)
- package depends on (purple arrow)
- package accesses (blue arrow)
- database populates (orange arrow)

Nodes:

- TNRS
- TNRS_pkg
- WorldFlora
- WorldFlora_pkg

Relationships:

- TNRS depends on TNRS_pkg
- TNRS depends on WorldFlora
- TNRS depends on WorldFlora_pkg
- TNRS_pkg depends on WorldFlora
- TNRS_pkg depends on WorldFlora_pkg
- WorldFlora depends on WorldFlora_pkg
- WorldFlora_pkg accesses WorldFlora
- WorldFlora_pkg populates WorldFlora

DÉMONSTRATION

A close-up photograph of a cluster of Lantana camara flowers, commonly known as Twin Lantana. The flowers are arranged in whorls and come in shades of pink, white, and yellow. A small, light-colored insect, possibly a crab spider, is visible on one of the flowers. The background is dark green foliage.

Focus sur les bases de plantes

Bases & outils spécifique pour les plantes

Identification de **13 bases & 19 paquets**

- 9 bases à l'échelle **mondiale**
- 4 **régionales** (FB 2020, GermanSL, USDA, VASCAN)
- Plusieurs paquets accèdent aux **même bases**

Relation entre bases ↔ paquets

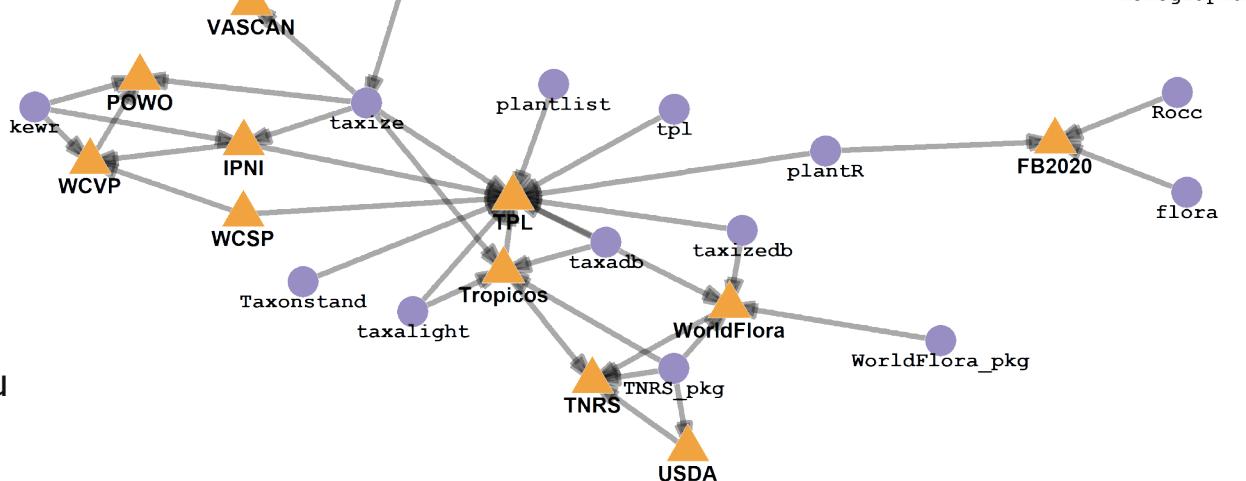
Centralité de **The Plant List**



Relation compliquées
entre bases

WCSP – WCVP – POWO ?

Les **bases régionales**
ne sont **pas connectés** au
reste du réseau



Perspectives sur les outils taxinomiques pour les plantes

The Plant List (obsolète depuis 9 ans!)

→ Facilité d'accès (API, pageweb *scrappable*, pérennité!)

≥4 taxonomies mondiale à jour

Leipzig Catalogue of Vascular Plants, POWO, World Flora Online, WorldPlants.de (CoL)

Pas d'effort d'harmonisation mondiale (contrairement aux oiseaux)

Article *in prep.* de comparaison des taxonomies globales avec TPL
(Mené by David Schellenberger-Costa)

Importance de rendre **liens** entre bases **explicites**

Liens entre bases **mondiale & régionales peu clairs**



Harmonisation taxinomique en pratique

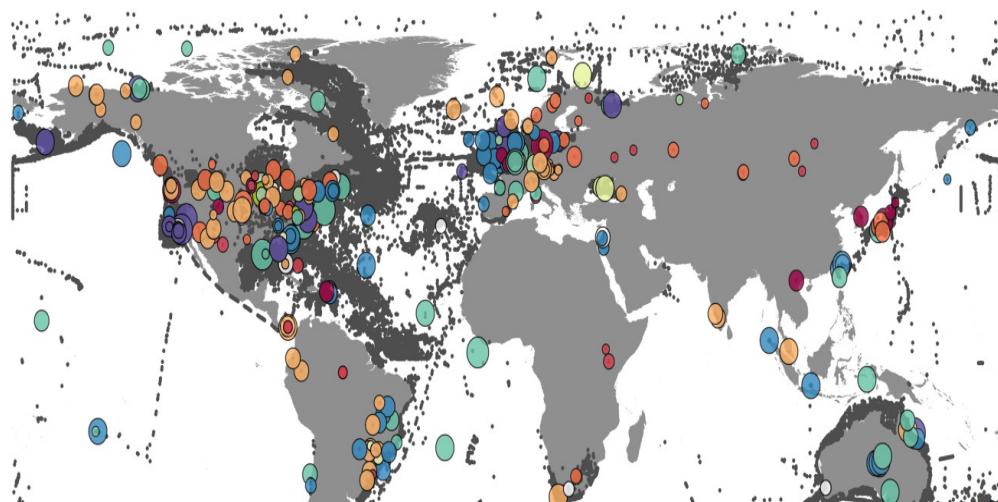
Cas d'utilisation

Comment harmoniser **en pratique** ?

Données
d'exemple



Dornelas et al. 2018



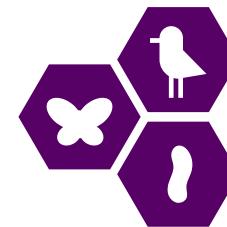
361 130 374 14 30
studies years contributors taxa biomes

Intérêt de BioTIME



Utilisée à iDiv by de nombreuses équipes
(+2000 téléchargements, >100 citations)

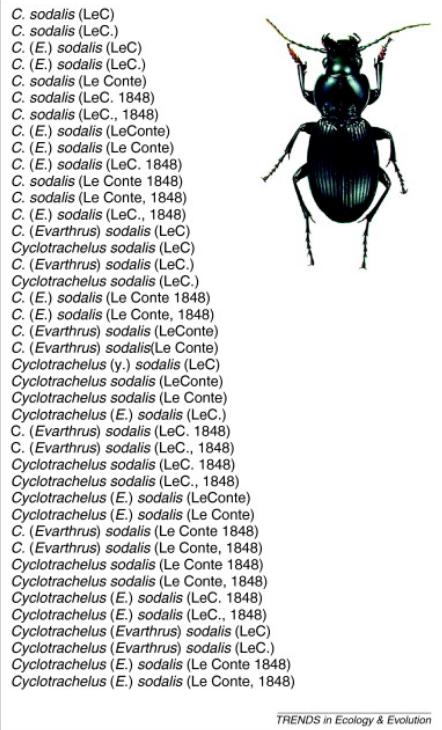
Multi-taxa → harmonisation plus difficile



Pas vraiment harmonisée

"Additionally, species names were checked for typographic errors and misspellings [...]. Most records were included as provided and may not always conform to the latest nomenclature." Dornelas et al. 2018

Importance du pré-traitement



Objectif : **Unifier le style d'écriture**
+ **corriger des erreurs**
(en identifiant les composants du noms)

***Cyclotrachelus sodalis* (Le Conte, 1848)**

44 326 noms
unique dans
BioTIME → **32 900** après **pré-traitement**

`rgnparser::gn_parse()`
`rgbif::parsenames()`

Patterson et al. 2010

Conclusion sur harmonisation en pratique

Analyse par **groupe taxonomique** réussissent **meilleures correspondances**

Analyse utilisant des **bases mondiales** sans restriction tax. sont **plus rapides** mais **résolvent moins de synonymes**

→ Le choix dépend de la question et du besoin en précision

A photograph of a wooden pier extending from the foreground into a body of water. The sky is filled with dramatic, colorful clouds at sunset. The word "Conclusion" is overlaid in large, white, sans-serif font in the center of the image.

Conclusion

Sortir du brouillard taxinomique



- **Penser à l'harmonisation taxonomique**
→ **nécessaire** pour maximiser la réutilisation de données
- **Connaître** le « paysage taxinomique »
→ utiliser **taxharmonizexplorer** 😎
- **Détailler ses analyses**
→ expliciter les paquets et fonctions utilisées

Perspectives

- Besoin de mieux mesurer l'impact **d'analyses différentes** sur les résultats (noms d'auteur, correspondance floue, etc.)
- Besoin de **dialogue** entre taxinomistes, gestionnaires de bases, et utilisateur·rice·s (écologues ou autres)
- Compléter le **réseau de ressources taxinomiques**





Merci !



taxharmonizexplorer

<https://tiny.cc/taxharmonizexplorer>