Tracing Disagreement in Taxonomy

Charles H. Pence • @pence@scholar.social

SPS, 2025-09-11

UCLouvain

isp PHILOSOPHIE

LIBST
LOUVAIN INSTITUTE OF BIOMOLECULAR SCIENCE AND TECHNOLOGY

ACADÉMIE ROYALE DE BELGIQUE 1772

Outline

1. Briefly: the problem
2. Construction of the corpus
3. Feature analysis
4. Preliminary results
5. Future ideas

Take-home message: Disagreement in taxonomy seems to be unevenly distributed; to understand it we'll have to analyze the literature empirically

The problem

Briefly: taxonomy is often taken to be affected by a widespread problem of **rampant uncertainty and disagreement**

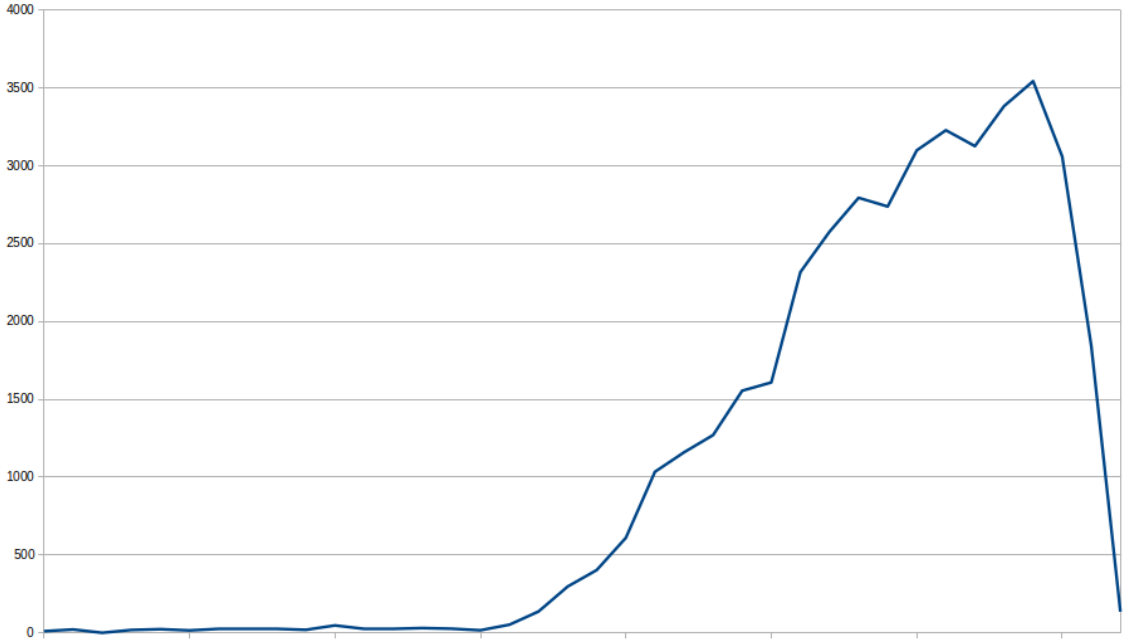
Part of the vast ornithology collection at the American Museum of Natural History.

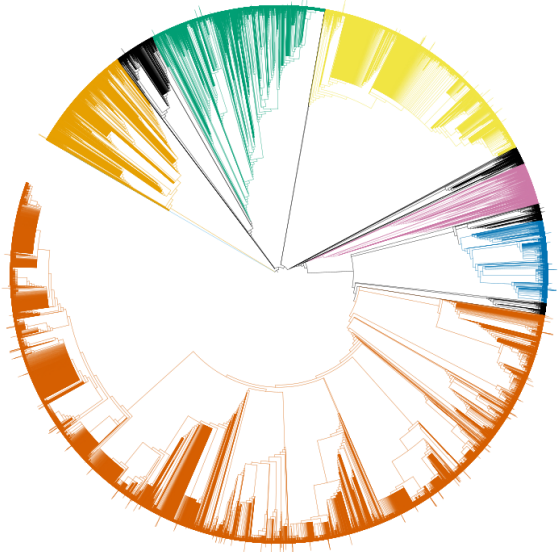
Taxonomy anarchy hampers conservation

The classification of complex organisms is in chaos.
Stephen T. Garnett and **Les Christidis** propose a solution.

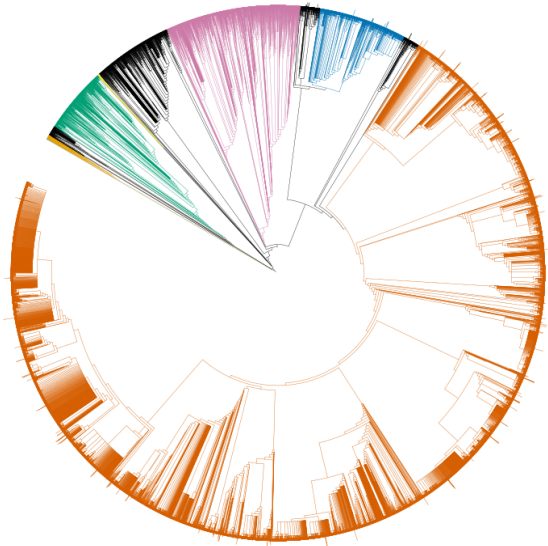
Corpus construction

Journal	Publisher	Size
Zootaxa	Magnolia Press	31,348
ZooKeys	Pensoft	4,940
PhytoKeys	Pensoft	820
Journal of Hymenoptera Research	Pensoft	382
MycoKeys	Pensoft	315
Zoosystematics and Evolution	Pensoft	153
Insecta Mundi	Center for Systematic Entomology	1,367
European Journal of Taxonomy	Museum National d'Histoire Naturelle	1,105


4000
3500
3000
2500
2000
1500
1000
500
0
1985
1990
1995
2000
2005
2010
2015
2020



Whole Open Tree of Life

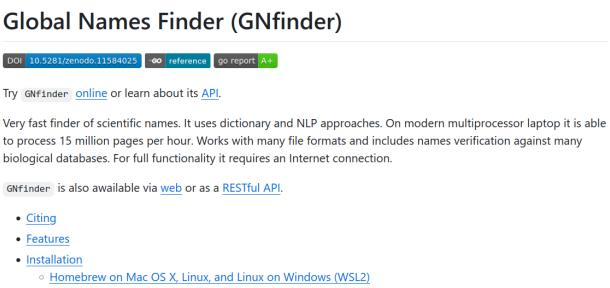


Corpus

Feature Analysis

Taxa

Global Names Finder (gnfinder): detect the names of species and other groups in text, both by comparison with global lists as well as detection of “probable” names



Global Names Finder (GNfinder)

DOI 10.5281/zenodo.11584025 | go reference | go report A+

Try `GNfinder` online or learn about its API.

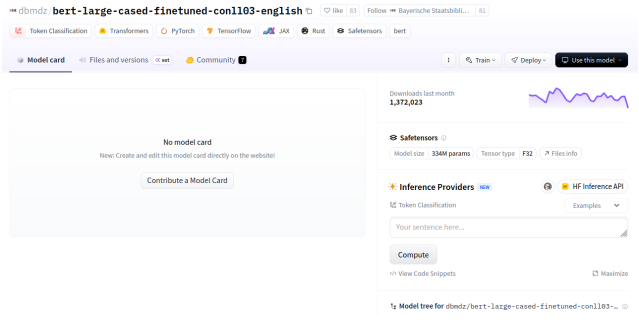
Very fast finder of scientific names. It uses dictionary and NLP approaches. On modern multiprocessor laptop it is able to process 15 million pages per hour. Works with many file formats and includes names verification against many biological databases. For full functionality it requires an Internet connection.

`GNfinder` is also awailable via web or as a RESTful API.

- Citing
- Features
- Installation
  - Homebrew on Mac OS X, Linux, and Linux on Windows (WSL2)

Locations

Pre-trained model for recognizing locations, organizations, and people in an English-language text (trained by the Bayerische Staatsbibliothek)



Topic Modeling

Convert documents into vectors in a 400-dimensional space (using the doc2vec algorithm), then examine clusters in this space. Normally, each cluster corresponds, more or less, to a subject of discussion.

Topic Modeling

But: **less useful than usual** in this corpus! Often, the clusters indicate how scientists talk about different groups of organisms ("fin, ray, gill, dorsal…"), though some might have a more interesting meaning ("barcoding, biodiversity, DNA…").

Disagreement

Close-reading of articles where we're sure that taxonomists are disagreeing with each other, to extract lists of keywords.

Désaccord

Example: the *disagreement* list

- critique
- doubt
- opinion
- disagree
- redundant
- reject
- rebuttal
- debate
- invalid
- misunderstanding
- misconception
- allegation
- allegedly
- mistake
- obsolete
- error
- misclassify
- erroneous
- contentious

Disagreement

3 lists: *epistemic values*, *disagreement*, and *pejorative evaluation*

Measure the relative frequency of these words in each article to give them a kind of “disagreement index.”

Methodology

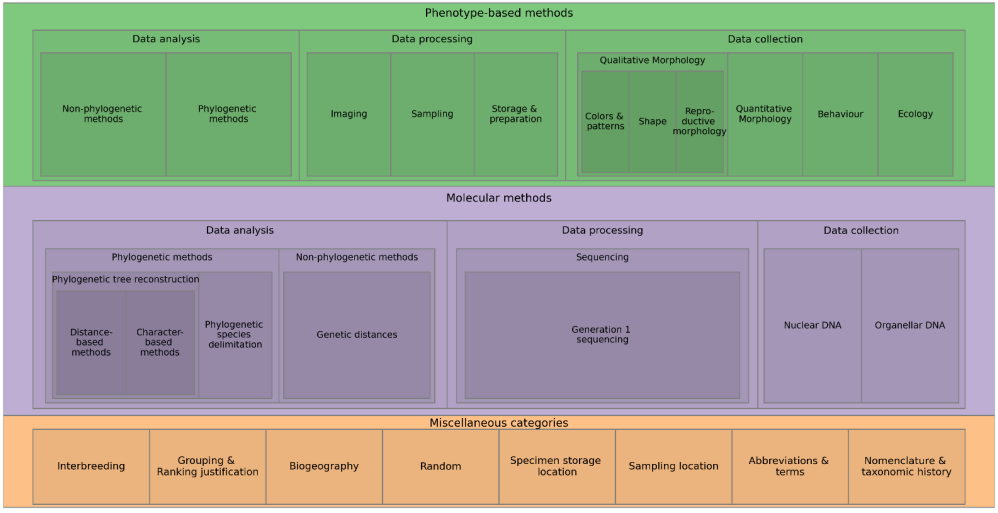
Difficult to detect in taxonomy:

- No “standard” citations for each method
- Different traditions of research (per taxon) = different terminology
- No tradition of describing your methods clearly (exploratory science)
- Lots of amateurs and researchers distributed across the world

Methodology

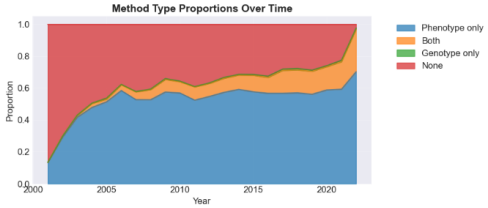
1. Lay out a general, hierarchical structure of methods
2. Isolate the “methods” sections
3. Exploratory analysis with topic modeling of these sections
4. Manual labeling of the paragraphs of these sections
5. Finalize the classification
6. Train classifiers/LLMs to classify the rest of the methods-paragraphs

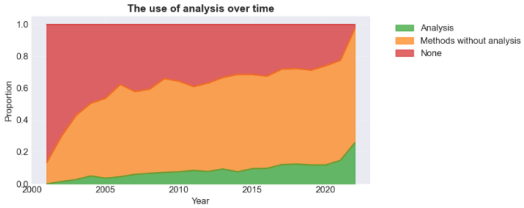
Methodology

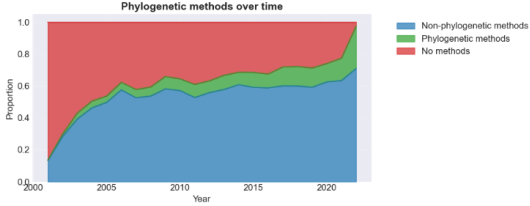


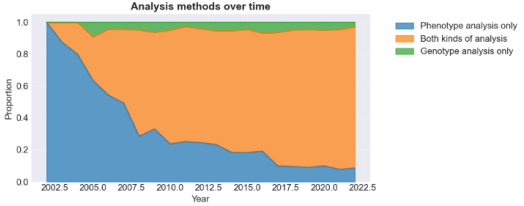
Preliminary results

Methodology

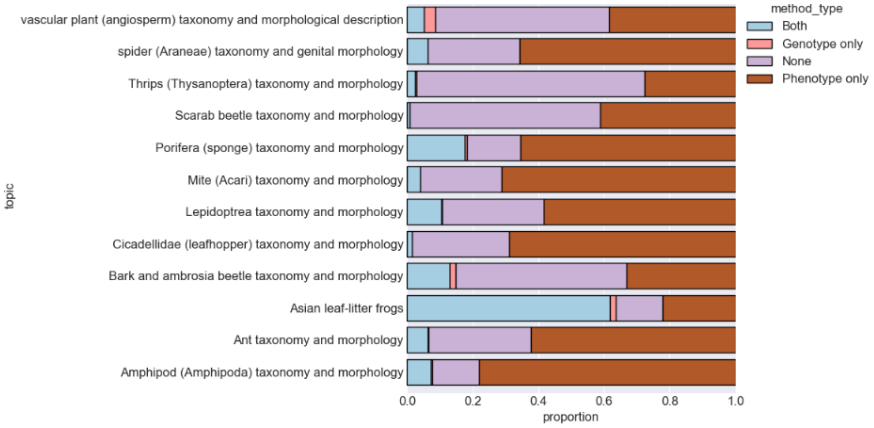




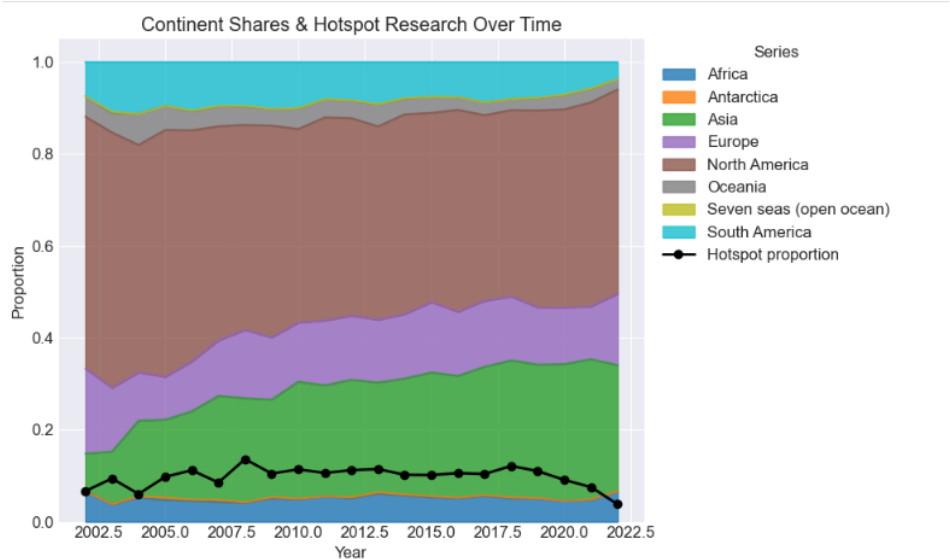




Methodology



Taxonomic attention



Disagreement and taxa

Divide organisms into “colloquial” groups (e.g., mammals, fish, birds, …).

- Lots more disagreement ($> 2\times$): birds ($n = 333$); mollusks ($n = 1064$)
- A bit more ($> 1.25\times$): mammals ($n = 396$)
- A bit less ($< 0.75\times$): fish ($n = 2132$); non-insect arthropods ($n = 7285$)
- A lot less ($< 0.5\times$): prokaryotes (! $n = 13$)

Disagreement and taxa

Second hypothesis: What about the **age** of the group? Test the correlation between the “disagreement index” and the year in which the main genus in the article was described.

We expect a **negative correlation**: the older the group, the more we argue about it.

Disagreement and taxa

Confirmed: **significant negative correlation**

An article on a genus described in 1750 should have a disagreement-index around 0.003 higher than one on a newly described genus (and 0.003 is around the mean disagreement index overall!).

Future ideas

Future ideas

- Correlations with places discussed (and especially eco-regions, biomes, etc.)
- In-depth analysis (close-reading) of changes in methodology with time and across taxa
- Construction of a “high-disagreement” corpus, then analysis of it to detect (maybe?) different senses/kinds of disagreement

Thanks to Stijn Conix, Tom Artois, Marlies Monnens, and Laura Vanstraelen!

Questions?

charles@charlespence.net • https://pencelab.be • @pence@scholar.social

PENCE LAB

fnrs
LA LIBERTÉ DE CHERCHER

Un problème avec les endroits

On ne peut faire de l'analyse des mots qui décrivent des endroits! Il faut les convertir en coordonnées latitude/longitude. Et faire cette conversion avec les outils normaux (*reverse geocoding*) est **très** cher.

Alors, utilisons un index géographique!

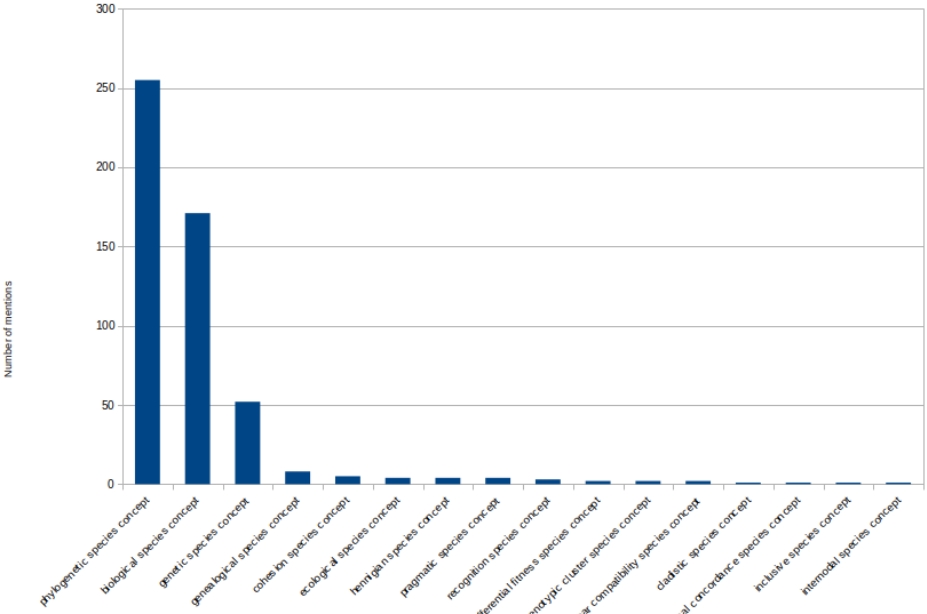
Index géographique

Extrêmement lent:

1. Téléchargez une liste des endroits et leurs coordonnées
2. Mettez le tout dans une base de données
3. Cherchez pour des résultats exacts
4. Sinon, cherchez pour des résultats approximatifs
5. S'il y en a plusieurs, calculez le «centre géographique» de l'article entier et prendre le résultat le plus proche

Concepts d'espèce

Phylo-Phenetic Species Concept
Phylogenetic Species Concept
Genic Species Concept
Cohesion Species Concept
Genealogical Concordance Species Concept
Genotypic Cluster Species Concept
Genetic Species Concept
Ecological Species Concept
Recognition Species Concept
Genealogical Species Concept
Biological Species Concept
Differential Fitness Species Concept
Compilospecies Concept
Cladistic Species Concept
Hennigian Species Concept
Internodal Species Concept
Mitonuclear Compatibility Species Concept
Pragmatic Species Concept
Inclusive Species Concept
Biosimilarity Species Concept


Number of mentions
300
250
200
150
100
50
0
phylogenetic species concept
biological species concept
genetic species concept
genealogical species concept
cohesion species concept
ecological species concept
hennigian species concept
pragmatic species concept
recognition species concept
differential fitness species concept
genotypic cluster species concept
nitonuclear compatibility species concept
cladistic species concept
genealogical concordance species concept
inclusive species concept
intermodal species concept