iDiv German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig

Taxonomic Harmonization

A problem you didn't know you had!

Matthias Grenié (matthias.grenie@idiv.de) – GfÖ & NFDI4Biodiversity Winter School
Lecture 8 – Thursday December 8th 2022

NFDI 4 BIODIVERSITY BIODIVERSITY, ECOLOGY & ENVIRONMENTAL DATA

GfÖ The Ecological Society of Germany, Austria and Switzerland

sDiv synthesis centre of iDiv

CC BY

iDiv is a research centre of the DFG Deutsche Forschungsgemeinschaft

"Museum of Natural History Cynthiacetus" by Shadowgate - CC BY 2.0

Who am I?

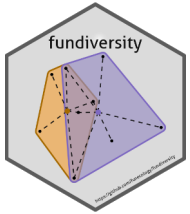
Who am I?

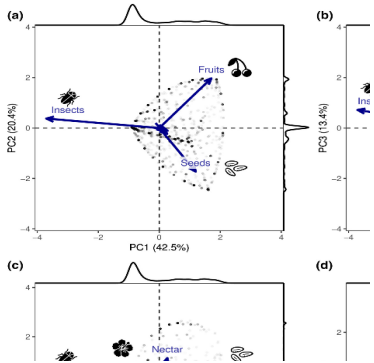
Postdoctoral Researcher at iDiv
Leipzig, Germany

Who am I?

Postdoctoral Researcher at iDiv
Leipzig, Germany

- Trait-based

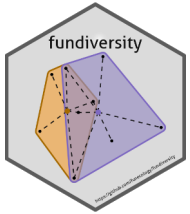


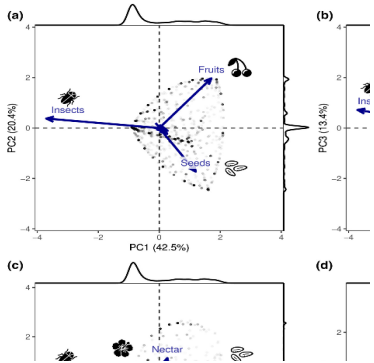


Who am I?

Postdoctoral Researcher at iDiv
Leipzig, Germany

- Trait-based





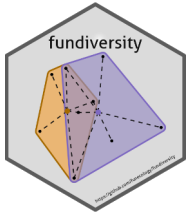
- Invasion

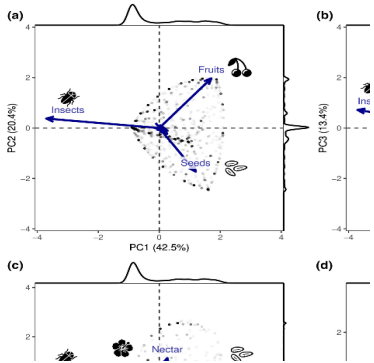
GloNAF
Global Naturalized Alien Flora

Who am I?

Postdoctoral Researcher at iDiv
Leipzig, Germany

- Trait-based

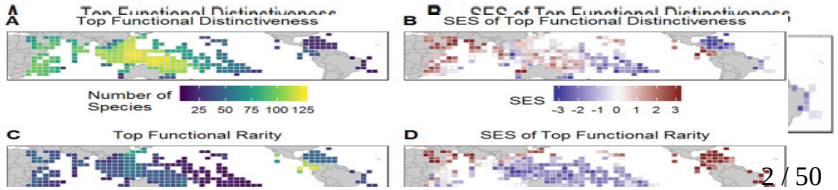




- Invasion

GloNAF
Global Naturalized Alien Flora

- Macro-ecologist



Contact details

- Email: matthias.grenie@idiv.de
- Twitter: @LeNematode
- Mastodon: @LeNematode@pouet.chapril.org
- Website: https://rekyt.github.io

Why am I in front of you here?

Why am I in front of you here?

Received: 7 September 2021 | Accepted: 21 December 2021

DOI: 10.1111/2041-210X.13802

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

REVIEW

Realising the Promise of Large Data and Complex Models

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

Matthias Grenié[1,2] | Emilio Berti[1,3] | Juan Carvajal-Quintero[1,2] | Gala Mona Louise Dädlow[1,2] | Alban Sagouis[1,4] | Marten Winter[1,2]

Programme of this lecture

Programme of this lecture

- What is taxonomy? (Taxonomy 101)

Programme of this lecture

- What is taxonomy? (Taxonomy 101)
- What is “taxonomic harmonization”?

Programme of this lecture

- What is taxonomy? (Taxonomy 101)
- What is “taxonomic harmonization”?
- What is needed to perform taxonomic harmonization?

Programme of this lecture

- What is taxonomy? (Taxonomy 101)
- What is “taxonomic harmonization”?
- What is needed to perform taxonomic harmonization?
- How to perform taxonomic harmonization in practice?

What is Taxonomy?

QUESTION
Who has a formal training in taxonomy?

"Hepaticae–Lebermoose" (1904) by Ernst Haeckel. Original from Library of Congress. Digitally enhanced by rawpixel." CC-BY 2.0.

Taxonomy is the basis of all ecology

"Hepaticae–Lebermoose" (1904) by Ernst Haeckel. Original from Library of Congress. Digitally enhanced by rawpixel." CC-BY 2.0.

"Hepaticae–Lebermoose" (1904) by Ernst Haeckel. Original from Library of Congress. Digitally enhanced by rawpixel." CC-BY 2.0.

Taxonomy is the basis of all ecology

Need to name things

"Hepaticae–Lebermoose" (1904) by Ernst Haeckel. Original from Library of Congress. Digitally enhanced by rawpixel." CC-BY 2.0.

Taxonomy is the basis of all ecology

Need to name things

Need to define entities we're working on

"Hepaticae–Lebermoose" (1904) by Ernst Haeckel. Original from Library of Congress. Digitally enhanced by rawpixel." CC-BY 2.0.

Taxonomy is the basis of all ecology

Need to name things

Need to define entities we're working on

Need to uncover (evolutionary) relationships between organisms

"Hepaticae–Lebermoose" (1904) by Ernst Haeckel. Original from Library of Congress. Digitally enhanced by rawpixel." CC-BY 2.0.

Taxonomy is the basis of all ecology

Need to name things

Need to define entities we're working on

Need to uncover (evolutionary) relationships between organisms

Historically taxonomy came before

Why should I care about taxonomy?

Why should I care about taxonomy?

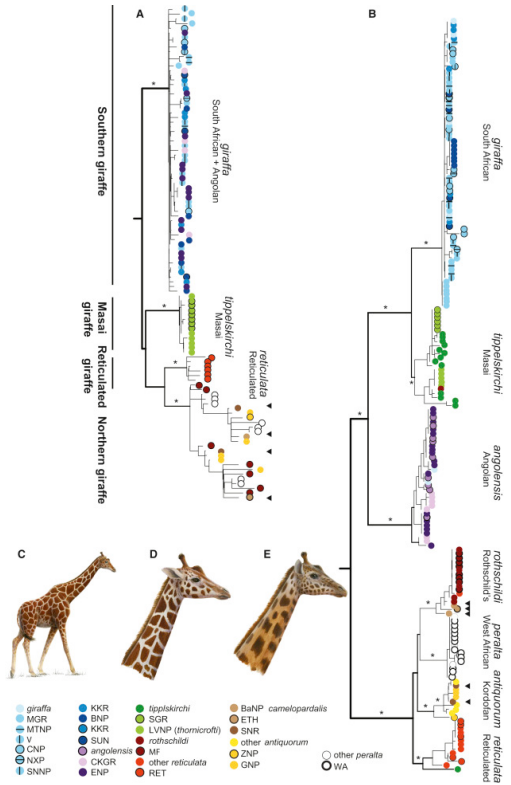
How many species of giraffe are there?

Why should I care about taxonomy?

How many species of giraffe are there?

Implications
for conservation programmes

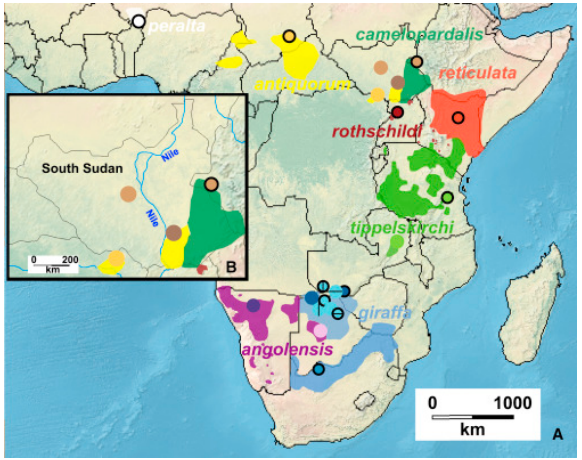
Why should I care about taxonomy?



How many species of giraffe are there?

↓

Implications
for conservation programmes

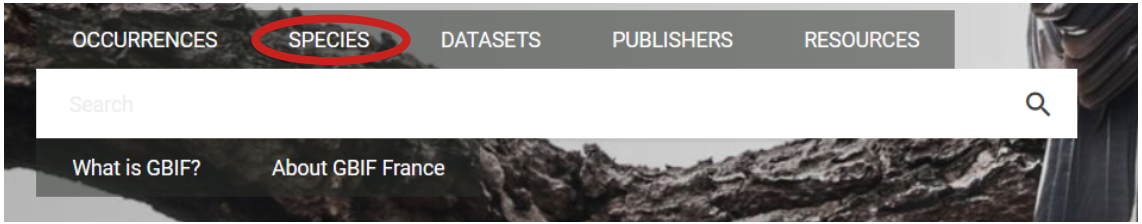


Fennessy et al. 2016 Current Biology

Everything is indexed by species names

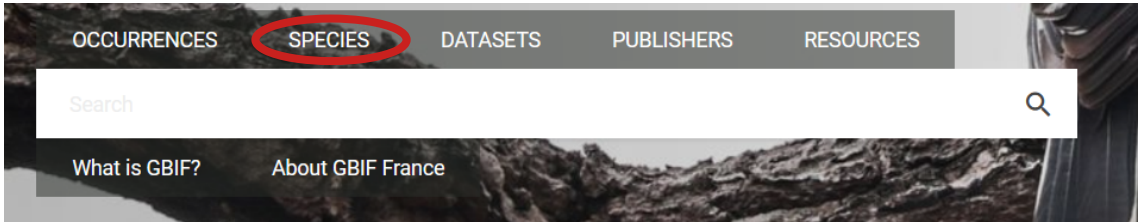
Everything is indexed by species names

Global Biodiversity Information Facility (GBIF) – Occurrence data



Everything is indexed by species names

Global Biodiversity Information Facility (GBIF) – Occurrence data

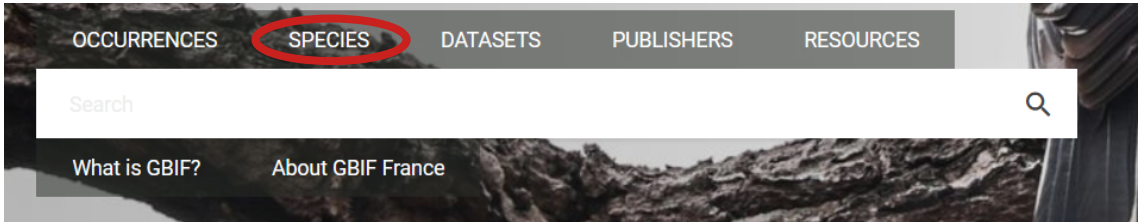


TRY Database – Plant Functional Trait

Recommended way to get data from TRY

Everything is indexed by species names

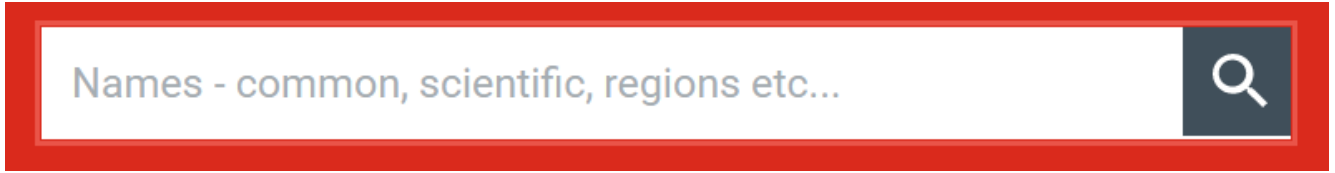
Global Biodiversity Information Facility (GBIF) – Occurrence data



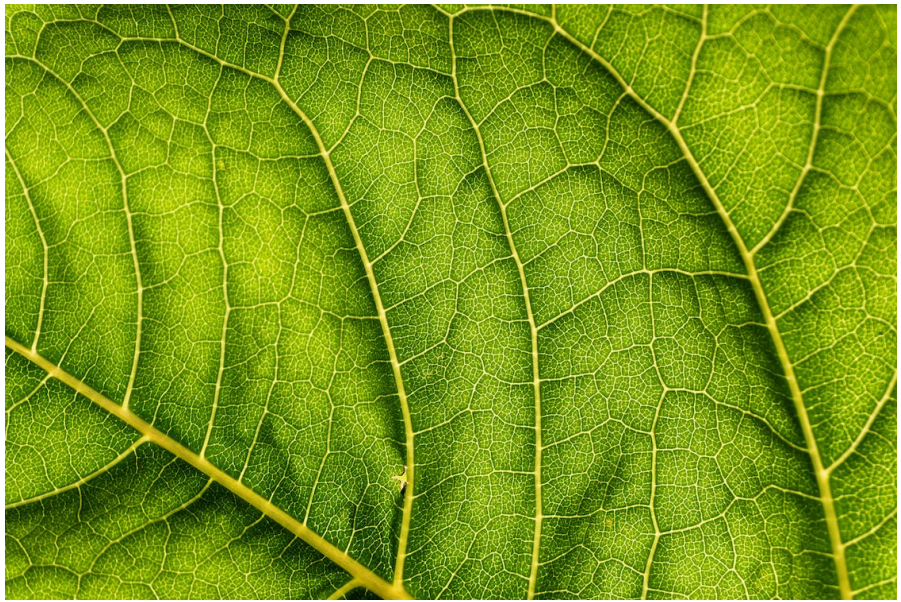
TRY Database – Plant Functional Trait

Recommended way to get data from TRY

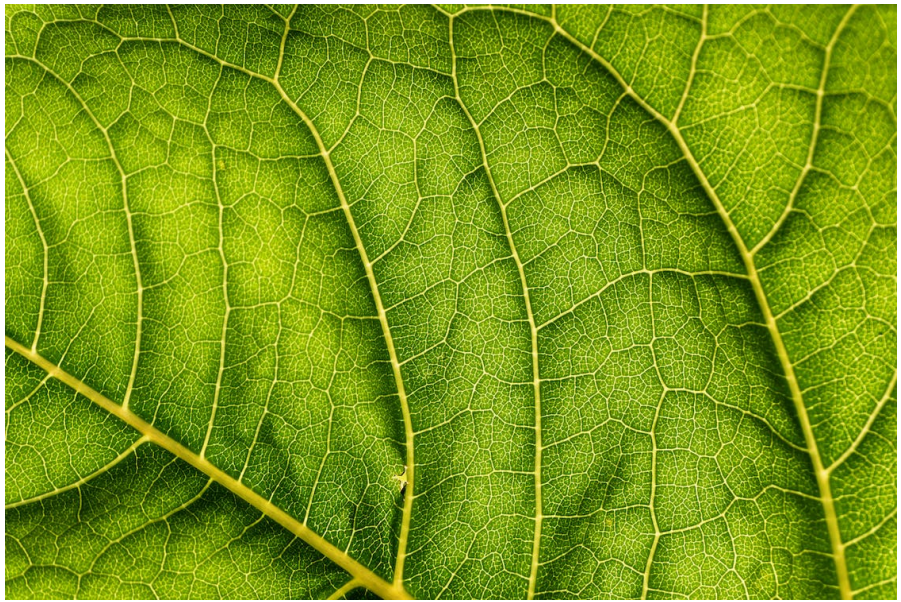
IUCN RedList – Threat Status



Taxonomy 101: how does it work?

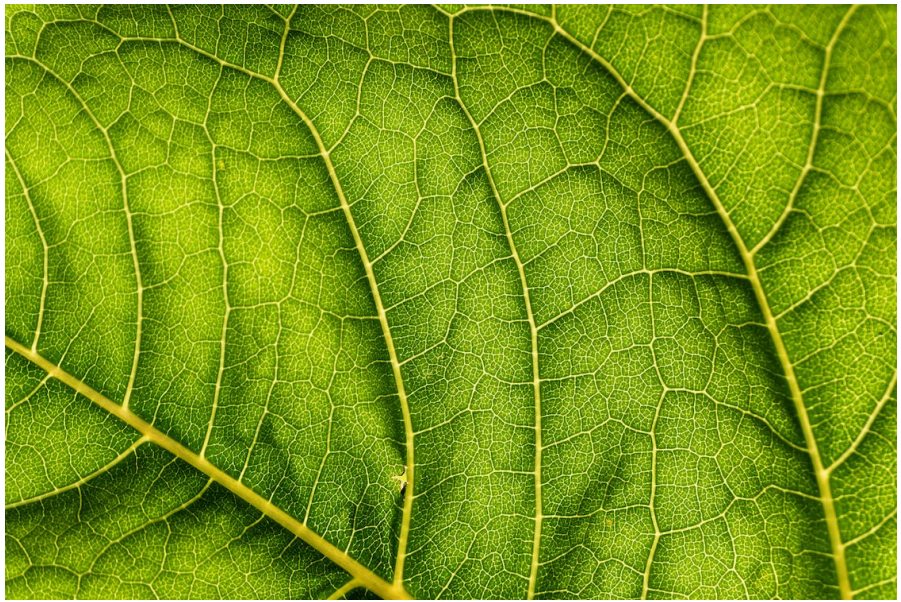


Taxonomy 101: how does it work?



Taxonomists use
Multiple Approaches
(DNA, Morphology, Behavior, etc.)

Taxonomy 101: how does it work?

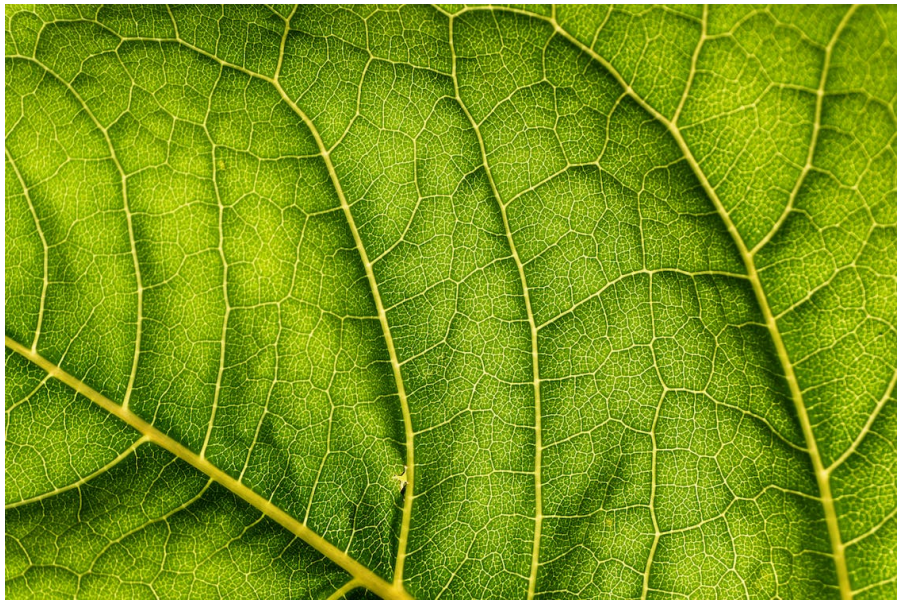


Taxonomists use
Multiple Approaches
(DNA, Morphology, Behavior, etc.)

↓

Establish
new species or relationships

Taxonomy 101: how does it work?



Taxonomists use
Multiple Approaches
(DNA, Morphology, Behavior, etc.)

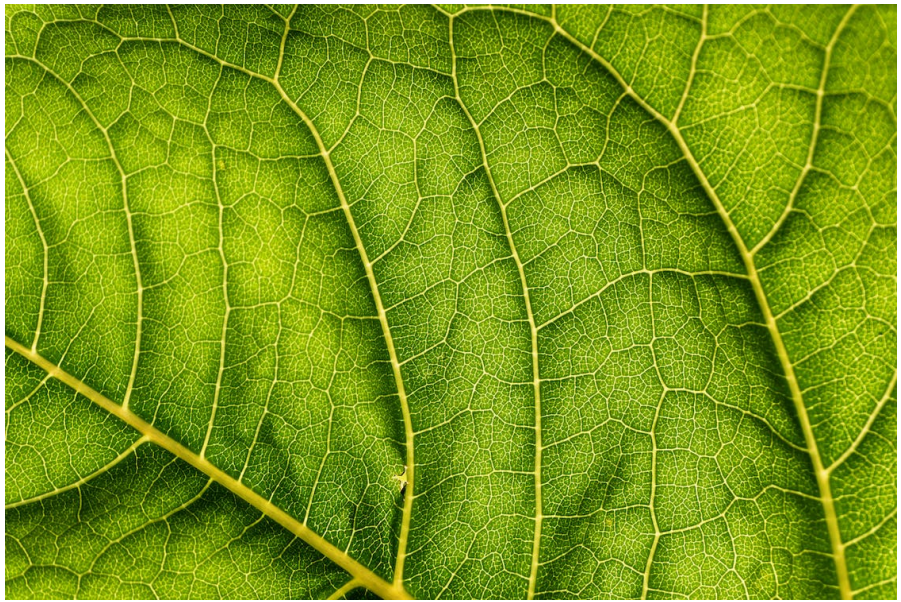
↓

Establish
new species or relationships

↓

Peer-reviewed into
specialist journals

Taxonomy 101: how does it work?



Taxonomists use
Multiple Approaches
(DNA, Morphology, Behavior, etc.)

↓

Establish
new species or relationships

↓

Peer-reviewed into
specialist journals

↓

Discussion and
Consensus Building

Aggregating taxonomic knowledge

Aggregating taxonomic knowledge

Taxonomic Information is **scattered**

Aggregating taxonomic knowledge

Taxonomic Information is **scattered**

Need to assemble them

Aggregating taxonomic knowledge

Taxonomic Information is **scattered**

↓

Need to assemble them

↓

Taxonomic Databases

Catalogue of Life

WoRMS
World Register of Marine Species

Aggregating taxonomic knowledge

Taxonomic Information is **scattered**

↓

Need to assemble them

↓

Taxonomic Databases

Catalogue of Life

WoRMS
World Register of Marine Species

Provide Up-to-date taxonomy

Rely on individual or community curation

Summary of this part

Summary of this part

- **Taxonomy** is at the **basis** of ecology and biodiversity sciences

Summary of this part

- **Taxonomy** is at the **basis** of ecology and biodiversity sciences
- Its main goal is to **discover** and **distinguish** the different **species**

Summary of this part

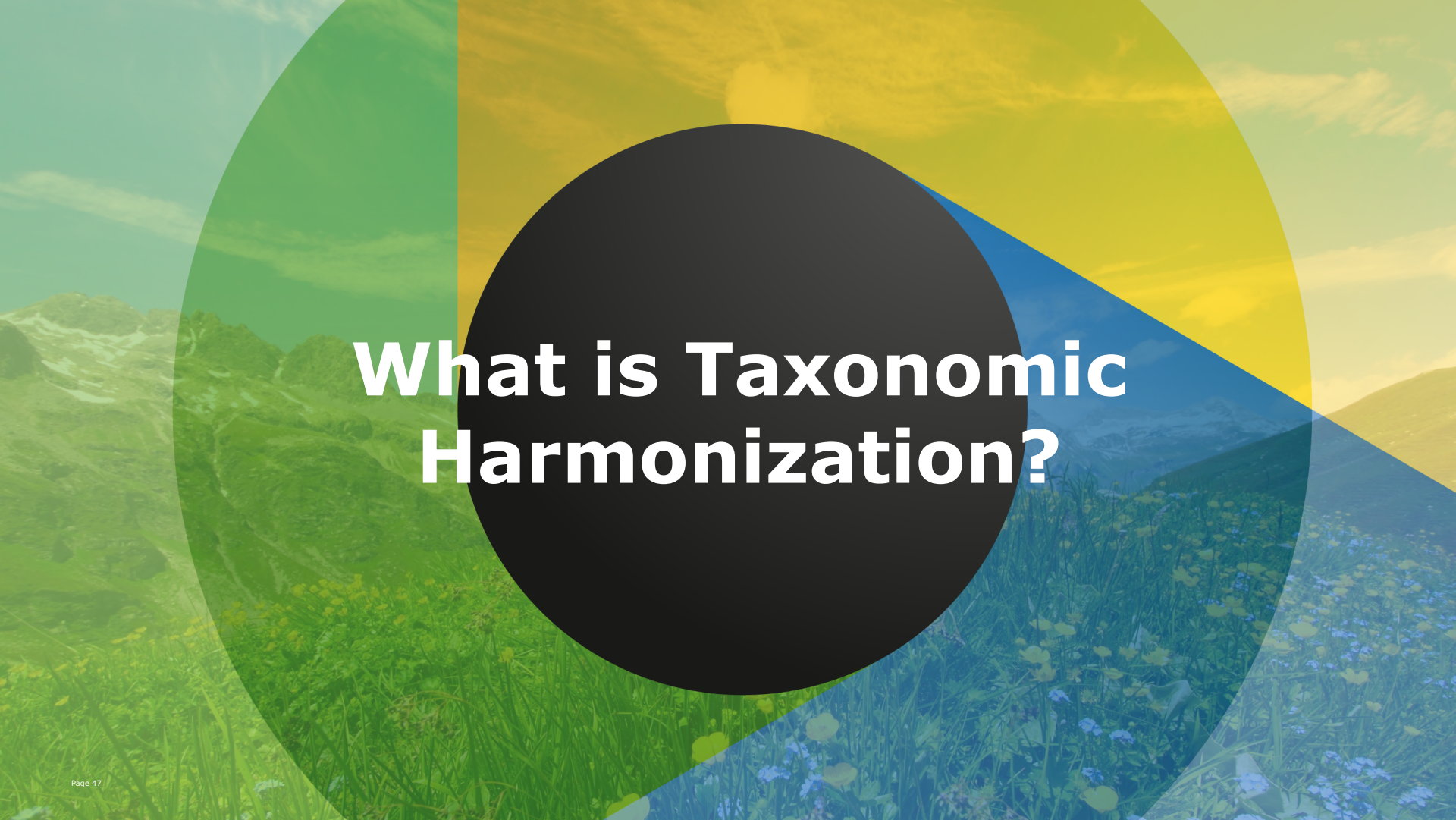
- **Taxonomy** is at the **basis** of ecology and biodiversity sciences
- Its main goal is to **discover** and **distinguish** the different **species**
- **Taxonomic information** (papers) is aggregated into **databases**

Summary of this part

- **Taxonomy** is at the **basis** of ecology and biodiversity sciences
- Its main goal is to **discover** and **distinguish** the different **species**
- **Taxonomic information** (papers) is aggregated into **databases**
- These databases are called **Taxonomic Reference Databases**

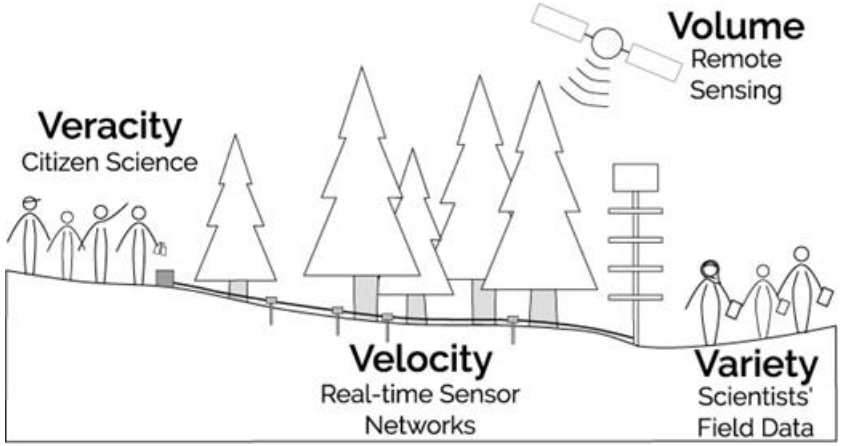
Summary of this part

- **Taxonomy** is at the **basis** of ecology and biodiversity sciences
- Its main goal is to **discover** and **distinguish** the different **species**
- **Taxonomic information** (papers) is aggregated into **databases**
- These databases are called **Taxonomic Reference Databases**
- They provide **updated taxonomic backbones** for other data



What is Taxonomic Harmonization?

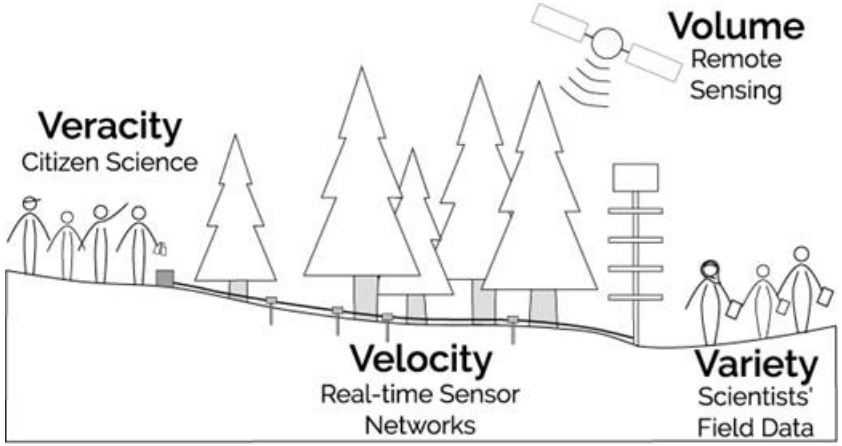
Increasing data size of Ecology



Farley et al. 2018 *BioScience*

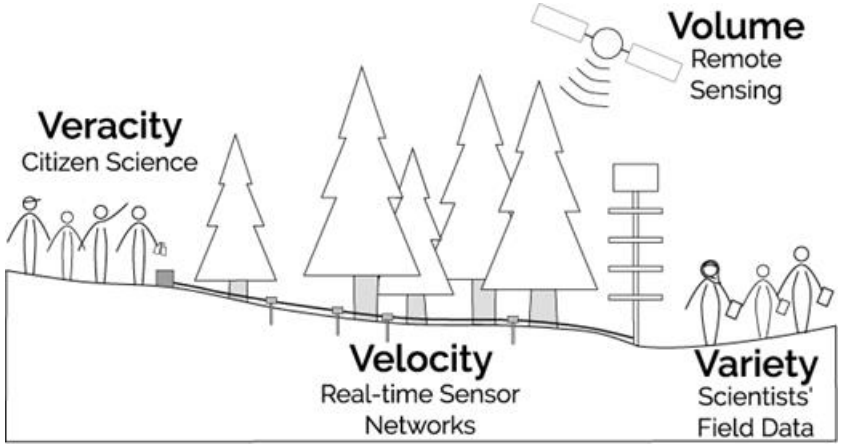
Increasing data size of Ecology

Data are getting larger



Farley et al. 2018 *BioScience*

Increasing data size of Ecology

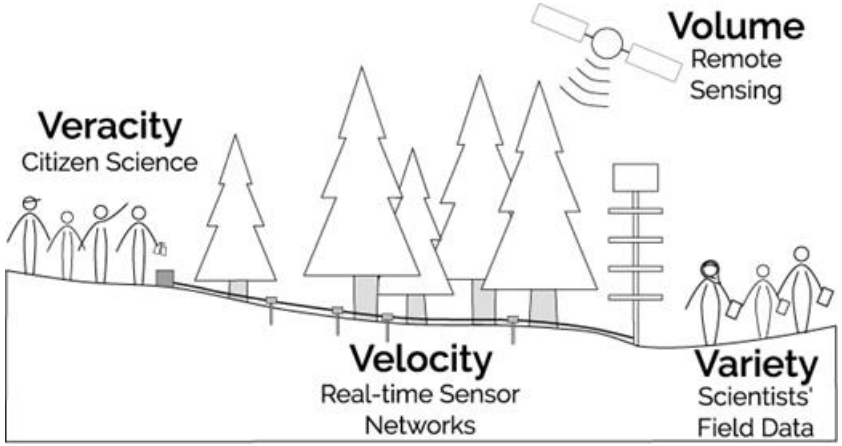


Farley et al. 2018 *BioScience*

Data are getting larger

Some aspects of “big data”

Increasing data size of Ecology



Farley et al. 2018 *BioScience*

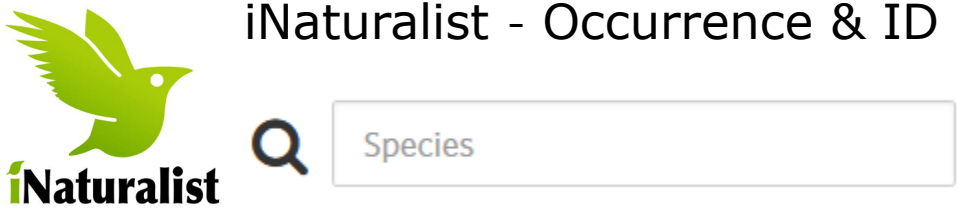
Data are getting larger

Some aspects of “big data”

Larger number of species measured

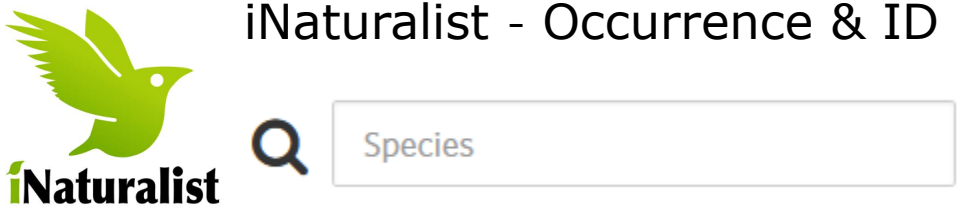
All data indexed by species name!

All data indexed by species name!

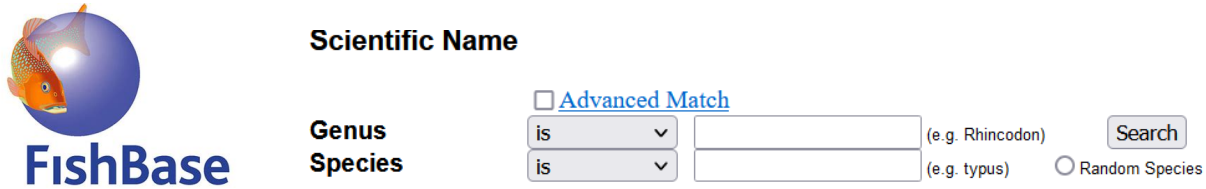


All data indexed by species name!

iNaturalist - Occurrence & ID

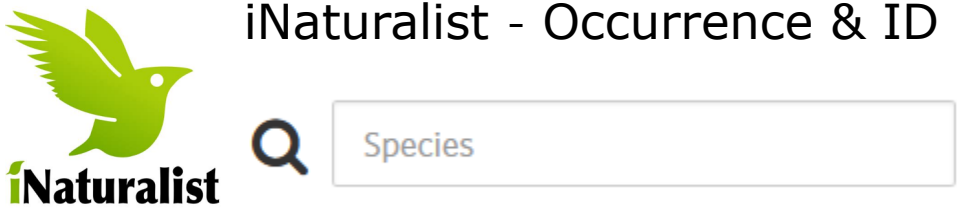


FishBase - Occurrence & Traits & Pictures & ...

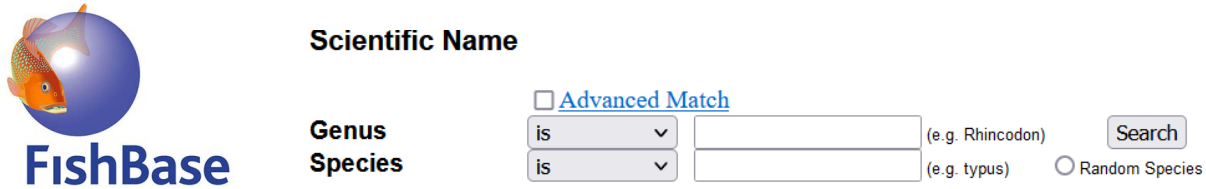


All data indexed by species name!

iNaturalist - Occurrence & ID



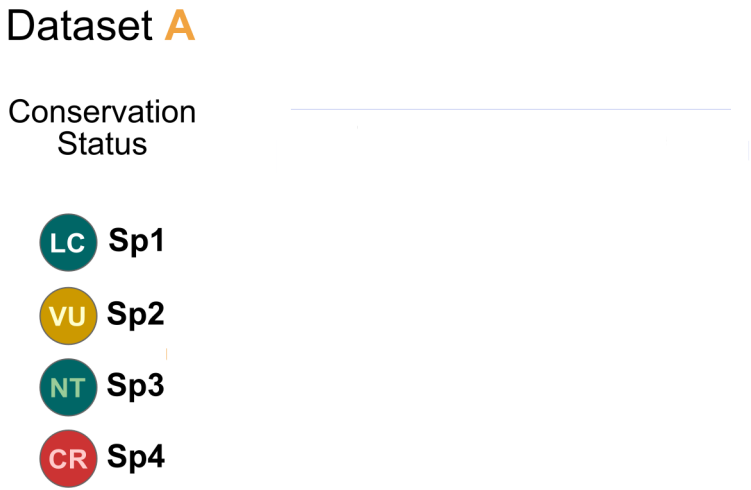
FishBase - Occurrence & Traits & Pictures & ...



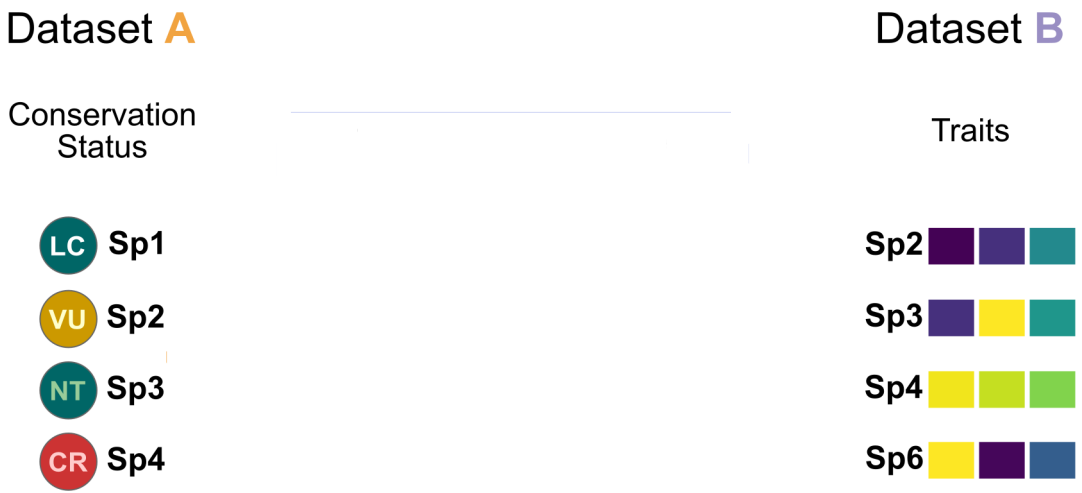
AlgaeBase - Distribution & Pictures & Taxonomy

The process of taxonomic harmonization

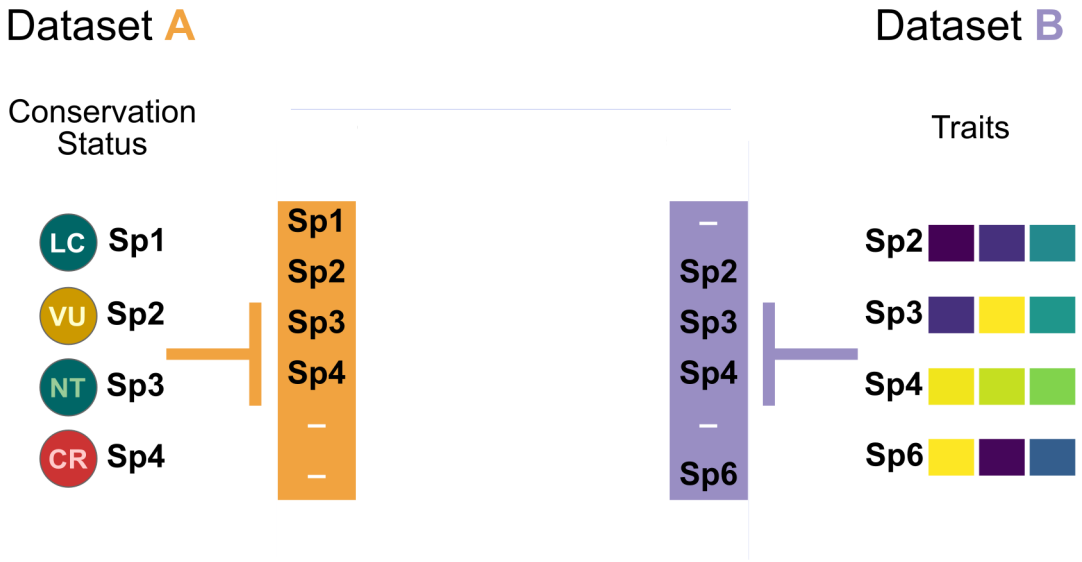
The process of taxonomic harmonization



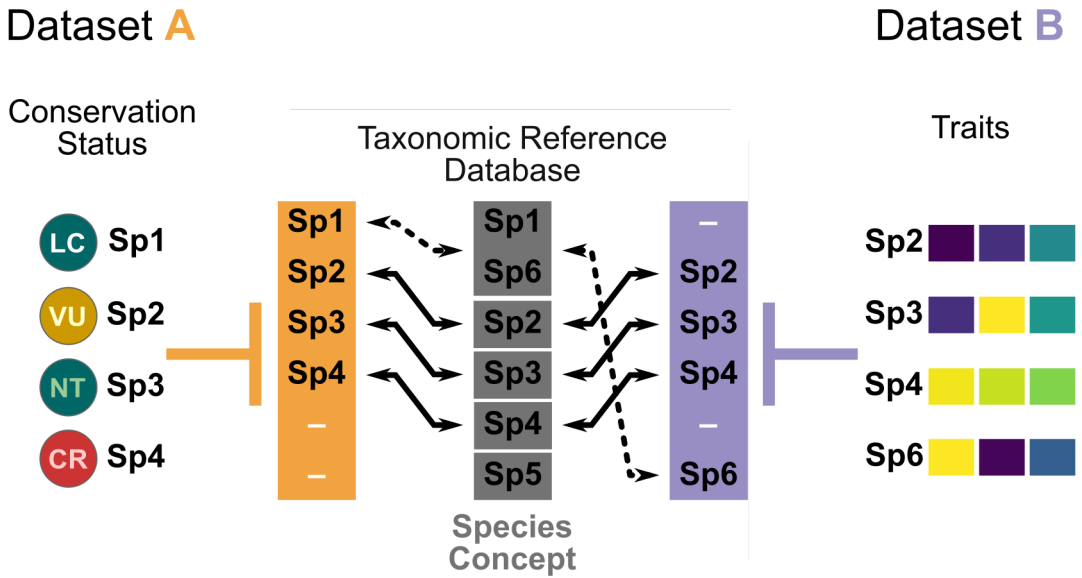
The process of taxonomic harmonization



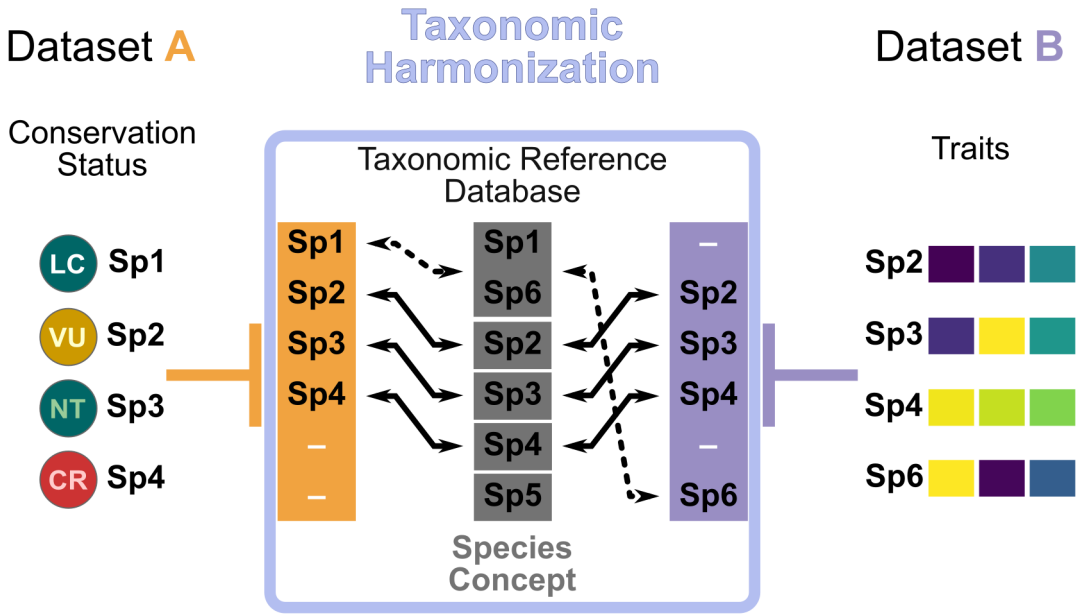
The process of taxonomic harmonization



The process of taxonomic harmonization



The process of taxonomic harmonization



A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020

A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020

Conservation Status
(IUCN)

Biomass
(EltonTraits)

A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020

Conservation Status
(IUCN)

Biomass
(EltonTraits)

Pipile pipile
Pipile cumanensis
Pipile cujubi
Pipile jacutinga
Megapodius decollatus
Margaroperdix madagarensis
Falcipennis falcipennis

A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020

Conservation Status
(IUCN)

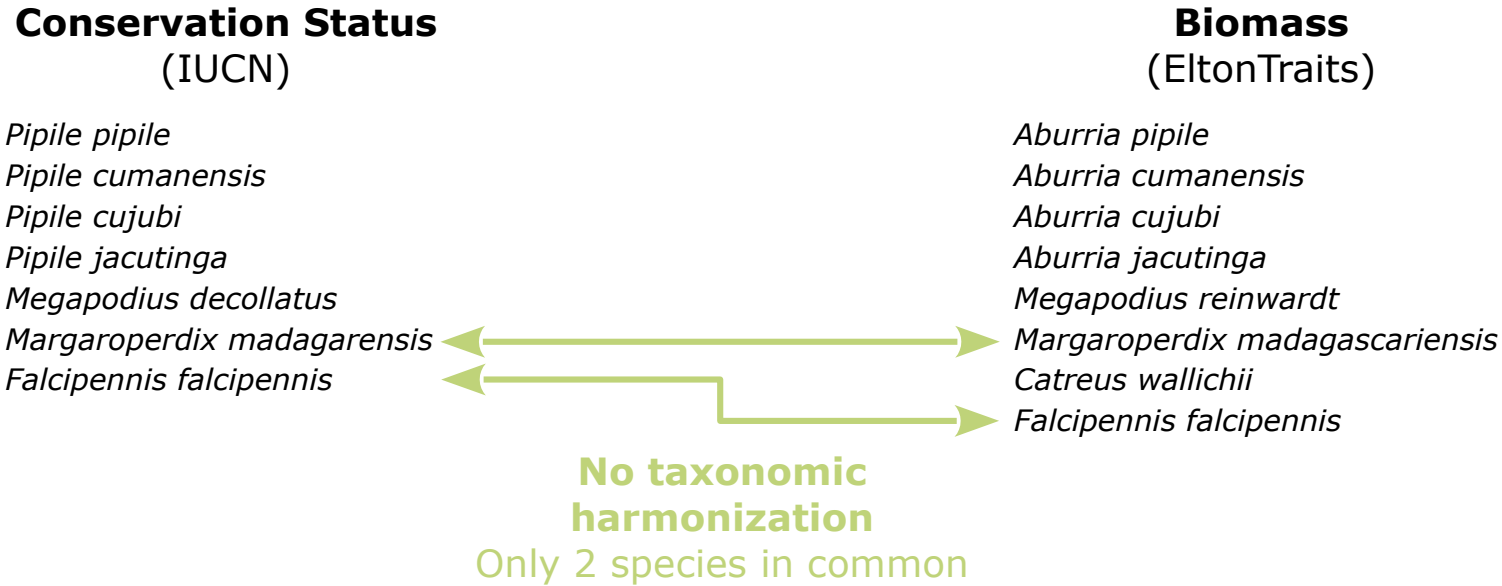
Pipile pipile
Pipile cumanensis
Pipile cujubi
Pipile jacutinga
Megapodius decollatus
Margaroperdix madagarensis
Falcipennis falcipennis

Biomass
(EltonTraits)

Aburria pipile
Aburria cumanensis
Aburria cujubi
Aburria jacutinga
Megapodius reinwardt
Margaroperdix madagascariensis
Catreus wallichii
Falcipennis falcipennis

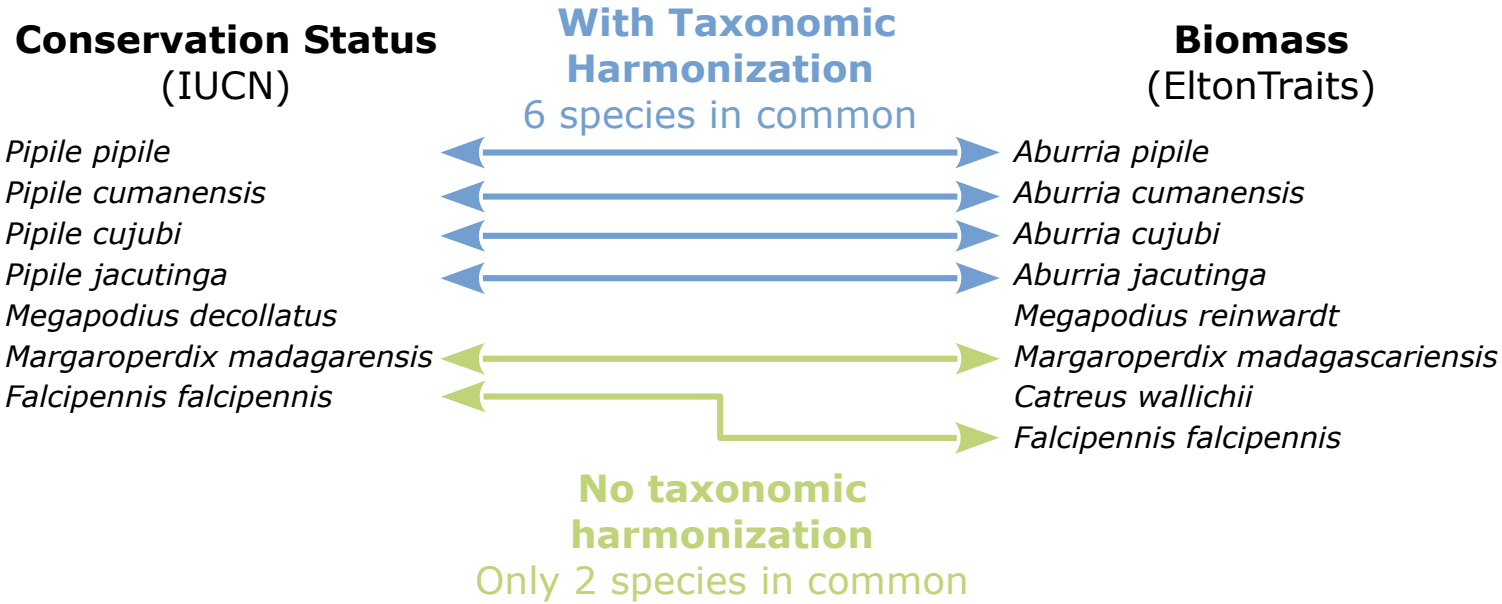
A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020



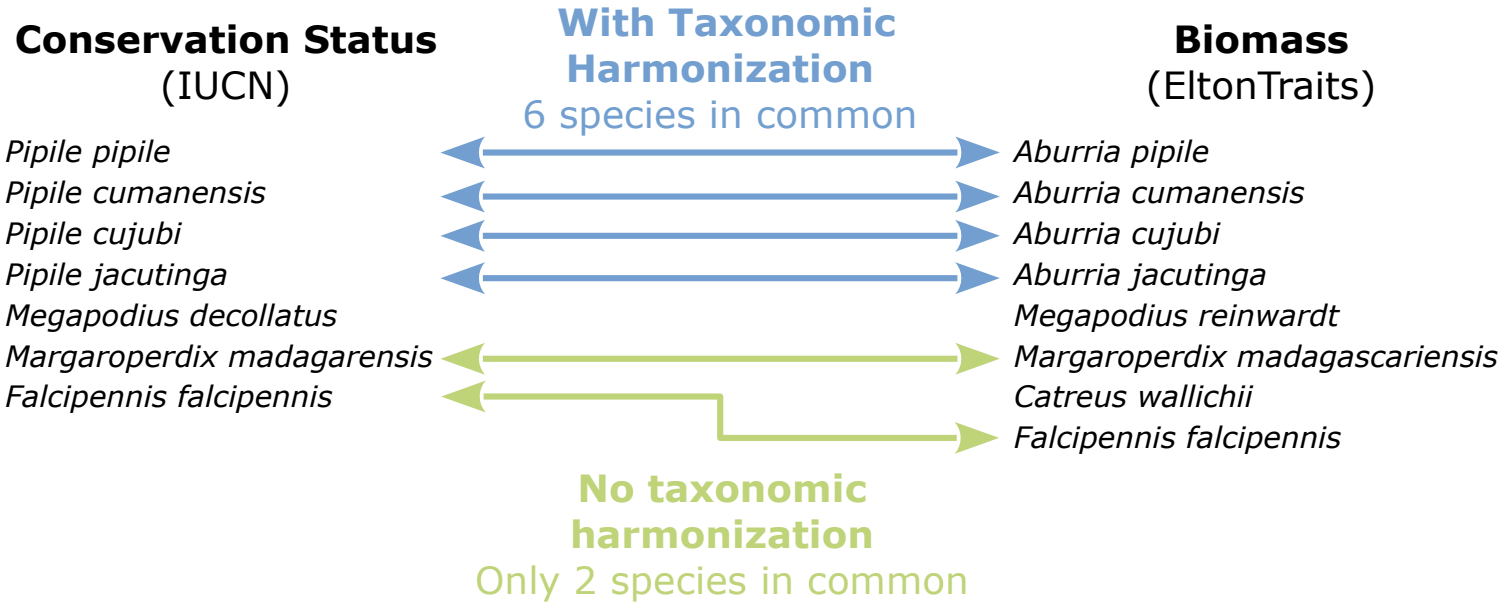
A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020



A concrete example: Bird Extinction Risk vs. Biomass

From Norman et al. 2020



Taxonomic harmonization make **datasets comparable**

QUESTION
Were you already aware of this issue?

What should you care?

"Insect drawer Natural History Museum Rose Stainthorp 2010"
by NHM Beetles and Bugs CC-BY 2.0.

What should you care?

Taxonomic Harmonization increases matched names

"Insect drawer Natural History Museum Rose Stainthorp 2010" by NHM Beetles and Bugs CC-BY 2.0.

What should you care?

Taxonomic Harmonization increases matched names

The taxonomy could have been updated between datasets

"Insect drawer Natural History Museum Rose Stainthorp 2010" by NHM Beetles and Bugs CC-BY 2.0.

What should you care?

Taxonomic Harmonization increases matched names

The taxonomy could have been updated between datasets

Use the last up-to-date taxonomy for your analyses

"Insect drawer Natural History Museum Rose Stainthorp 2010" by NHM Beetles and Bugs CC-BY 2.0.

Summary of this part

Summary of this part

- **Taxonomic Harmonization** is the matching of species names against species reference databases

Summary of this part

- **Taxonomic Harmonization** is the matching of species names against species reference databases
- Taxonomic Harmonization is **necessary** with the **increasing size** of datasets in ecology (manual harmonization is too hard)

Summary of this part

- **Taxonomic Harmonization** is the matching of species names against species reference databases
- Taxonomic Harmonization is **necessary** with the **increasing size** of datasets in ecology (manual harmonization is too hard)
- With taxonomic harmonization we can resolve **given species names** into updated ones and take care of **synonyms**

What is needed to perform Taxonomic Harmonization?

The sources: Taxonomic Reference Databases

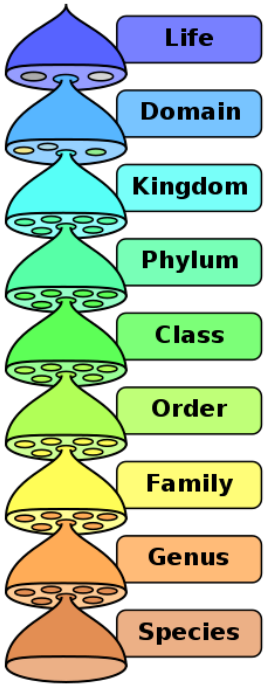
The sources: Taxonomic Reference Databases

Taxonomic Reference Database
(also taxonomic backbone, taxonomic checklist, taxonomic authority, etc.)

The sources: Taxonomic Reference Databases

Taxonomic Reference Database
(also taxonomic backbone, taxonomic checklist, taxonomic authority, etc.)

"Centralized repository of nomenclatural information"

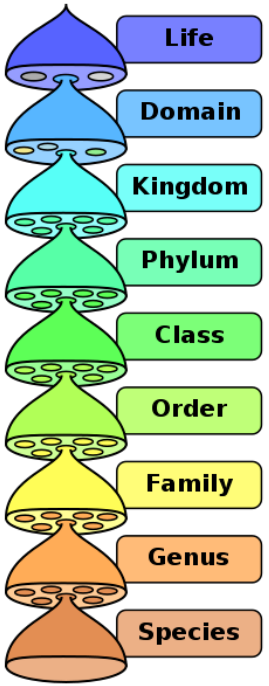


The sources: Taxonomic Reference Databases

Taxonomic Reference Database
(also taxonomic backbone, taxonomic checklist, taxonomic authority, etc.)

"Centralized repository of nomenclatural information"

+Spelling Correction **+Synonymy Resolution**



Examples of Taxonomic Reference Databases

Examples of Taxonomic Reference Databases

Backbone
GBIF (2021)

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

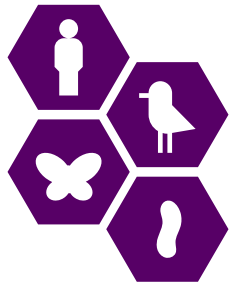
No taxonomic restriction

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global



No taxonomic restriction

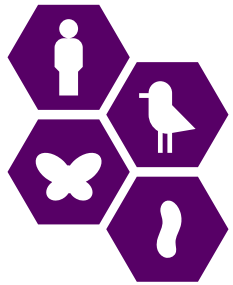
June 2021
6.6M names

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global



No taxonomic restriction

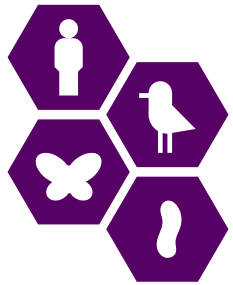
June 2021
6.6M names
3.7M accepted

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global



No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Leipzig **C**atalogue
of **V**ascular **P**lants
LCVP
(Freiberg et al. 2020)

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Leipzig **C**atalogue of **V**ascular **P**lants
LCVP
(Freiberg et al. 2020)

Global

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Leipzig **C**atalogue of **V**ascular **P**lants
LCVP
(Freiberg et al. 2020)

Global

Vascular Plants

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Leipzig **C**atalogue of **V**ascular **P**lants
LCVP
(Freiberg et al. 2020)

Global

Vascular Plants

Nov. 2020
1.3M names

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Leipzig **C**atalogue of **V**ascular **P**lants
LCVP
(Freiberg et al. 2020)

Global

Vascular Plants

Nov. 2020
1.3M names
350k accepted

Examples of Taxonomic Reference Databases

GBIF

Backbone
GBIF (2021)

Global

No taxonomic restriction

June 2021
6.6M names
3.7M accepted
2.6M synonyms

Leipzig **C**atalogue of **V**ascular **P**lants
LCVP
(Freiberg et al. 2020)

Global

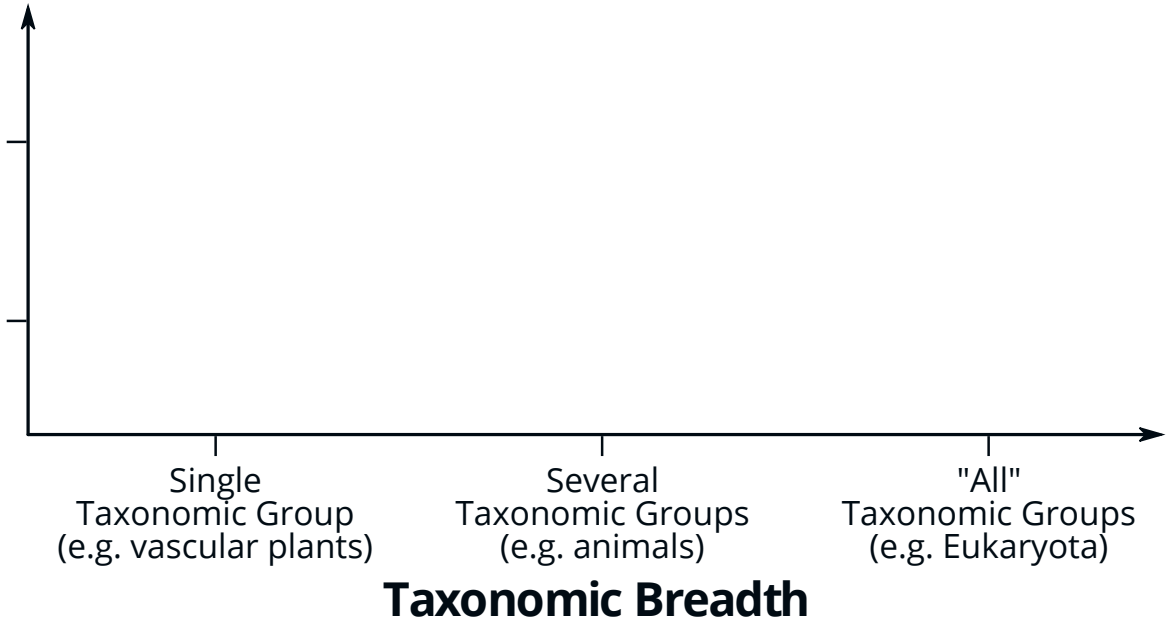
Vascular Plants

Nov. 2020
1.3M names
350k accepted
850k synonyms

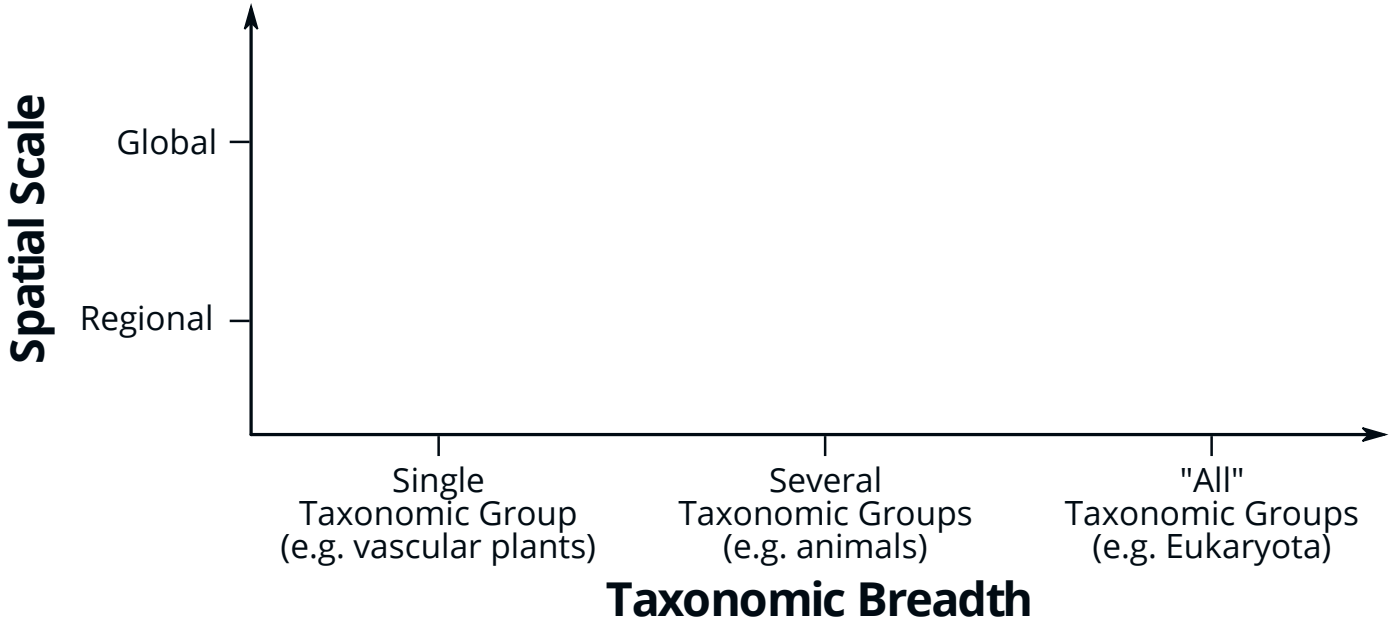
QUESTION
What taxonomic reference database do you know for your group of interest?
(Microbes excepted)

Database typology: Spatial Scale vs. Taxonomic Breadth

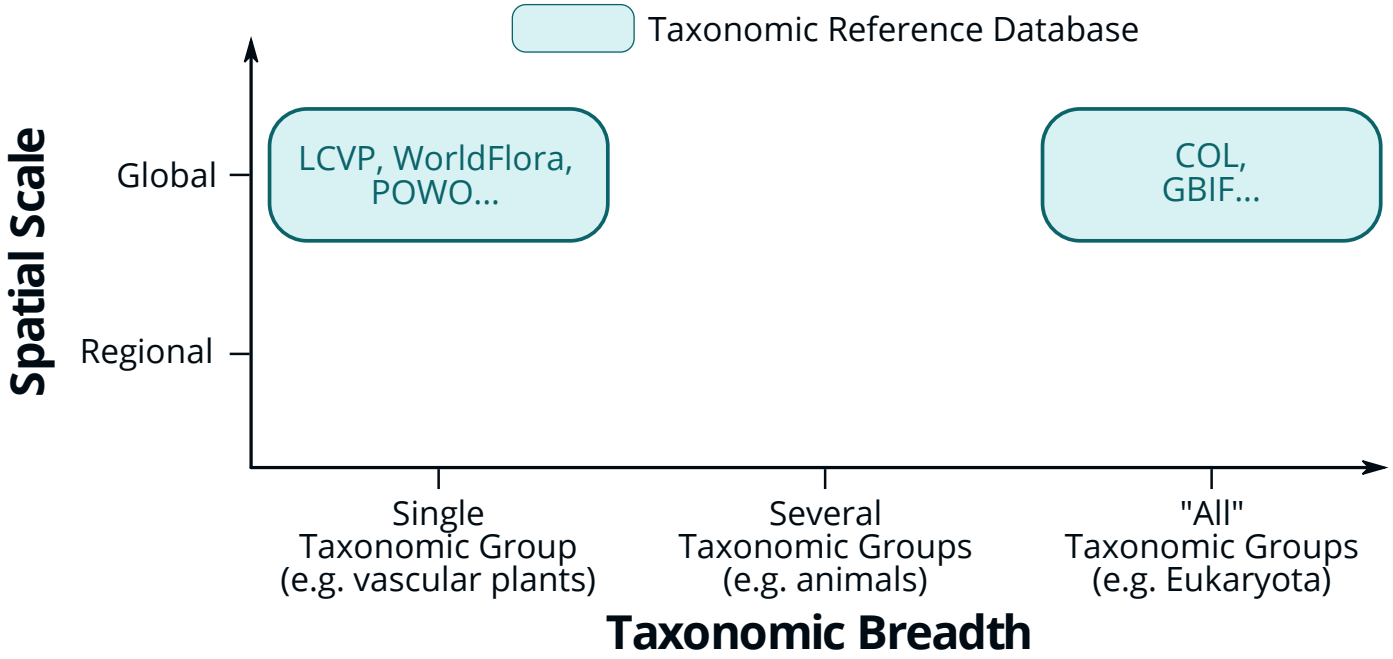
Database typology: Spatial Scale vs. Taxonomic Breadth



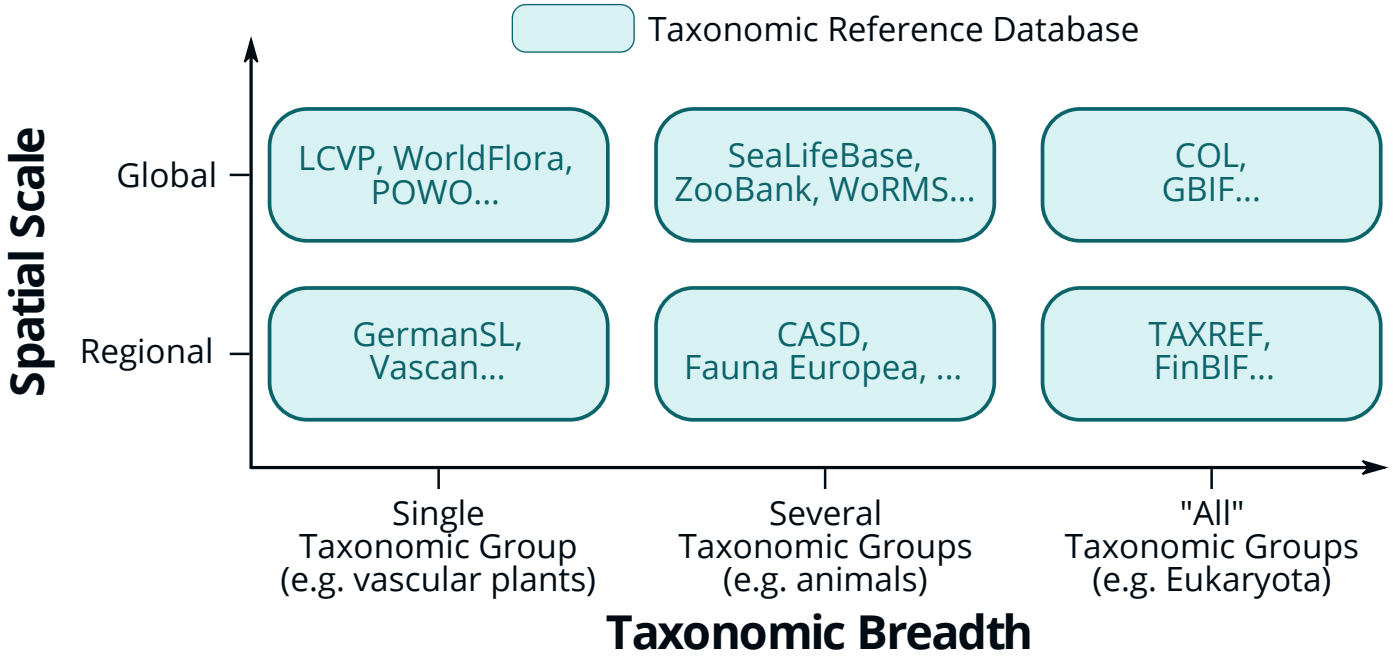
Database typology: Spatial Scale vs. Taxonomic Breadth



Database typology: Spatial Scale vs. Taxonomic Breadth

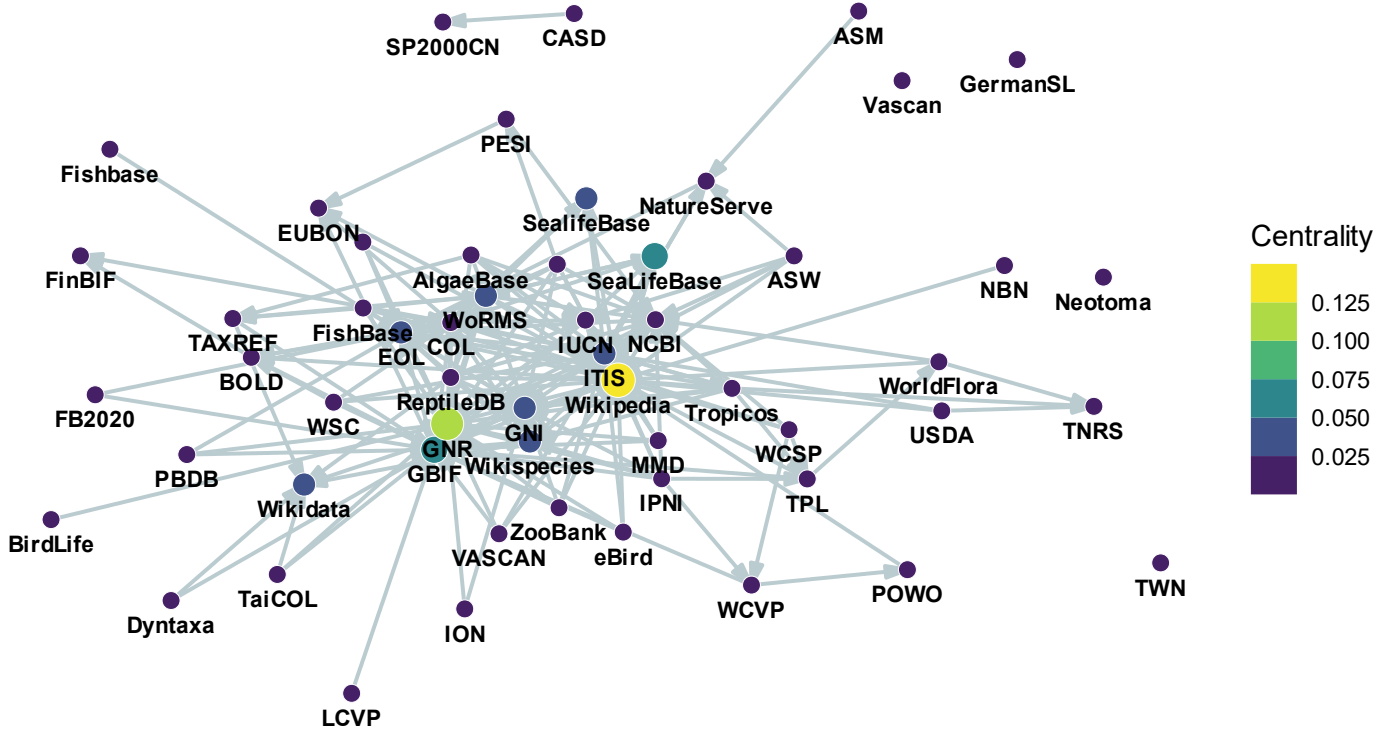


Database typology: Spatial Scale vs. Taxonomic Breadth



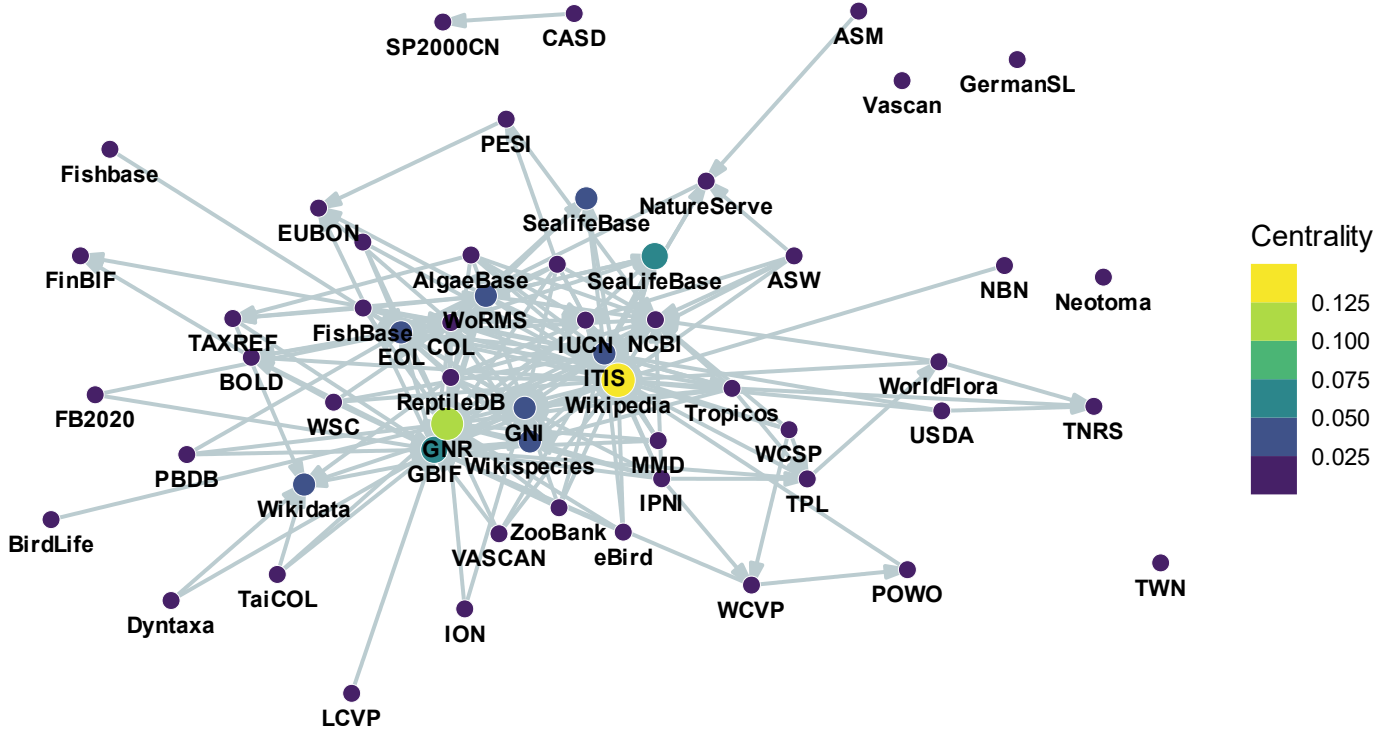
Beware of connections across databases

Beware of connections across databases



Beware of connections across databases

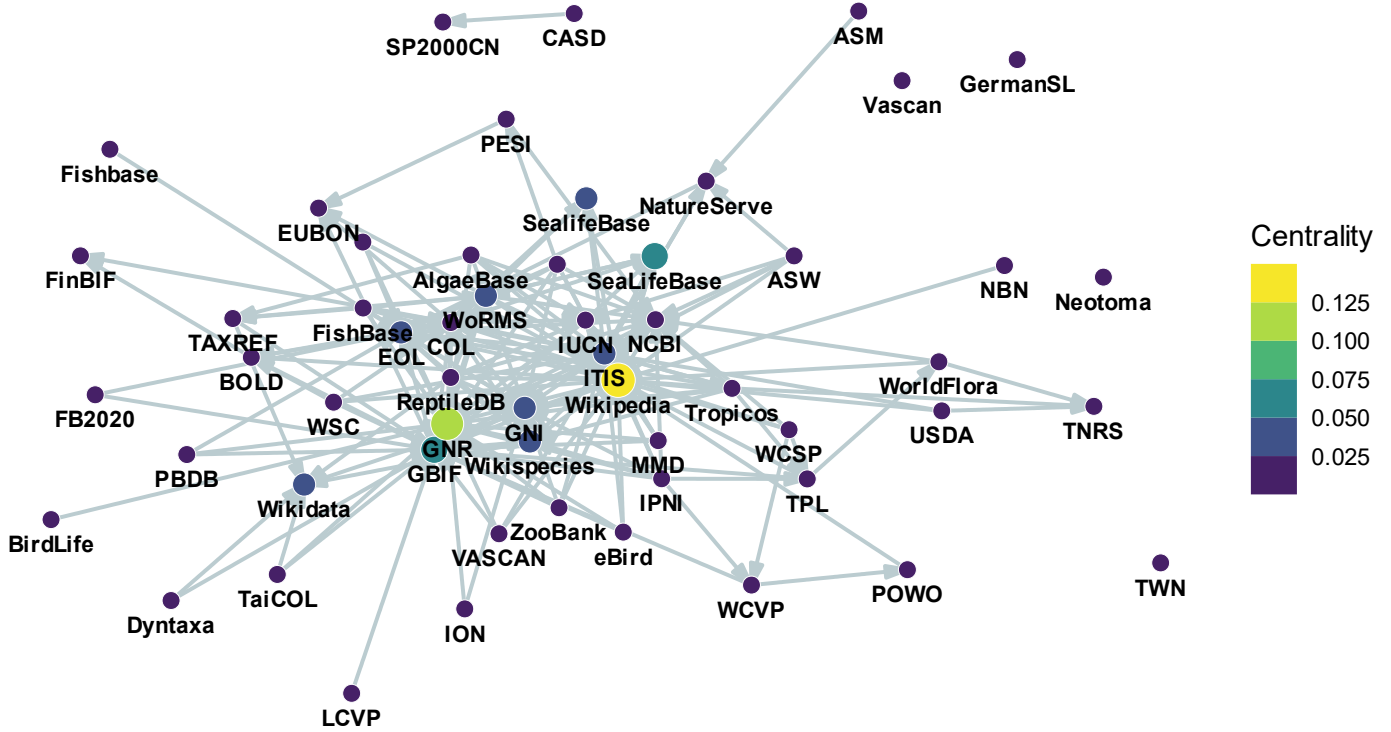
Relationships
Source → Target



Beware of connections across databases

Relationships
Source → Target

Specific databases
widely reused

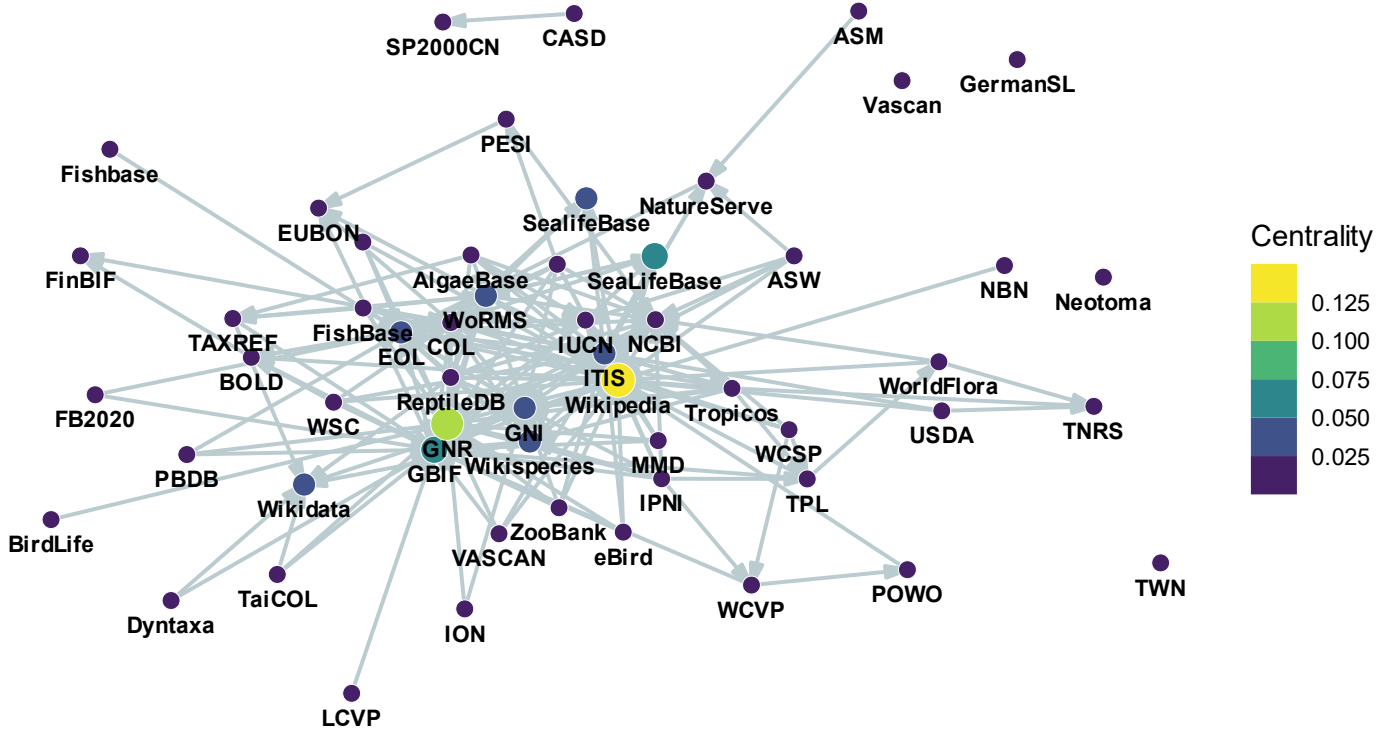


Beware of connections across databases

Relationships
Source → Target

Specific databases
widely reused

Some "mega-" aggregator databases

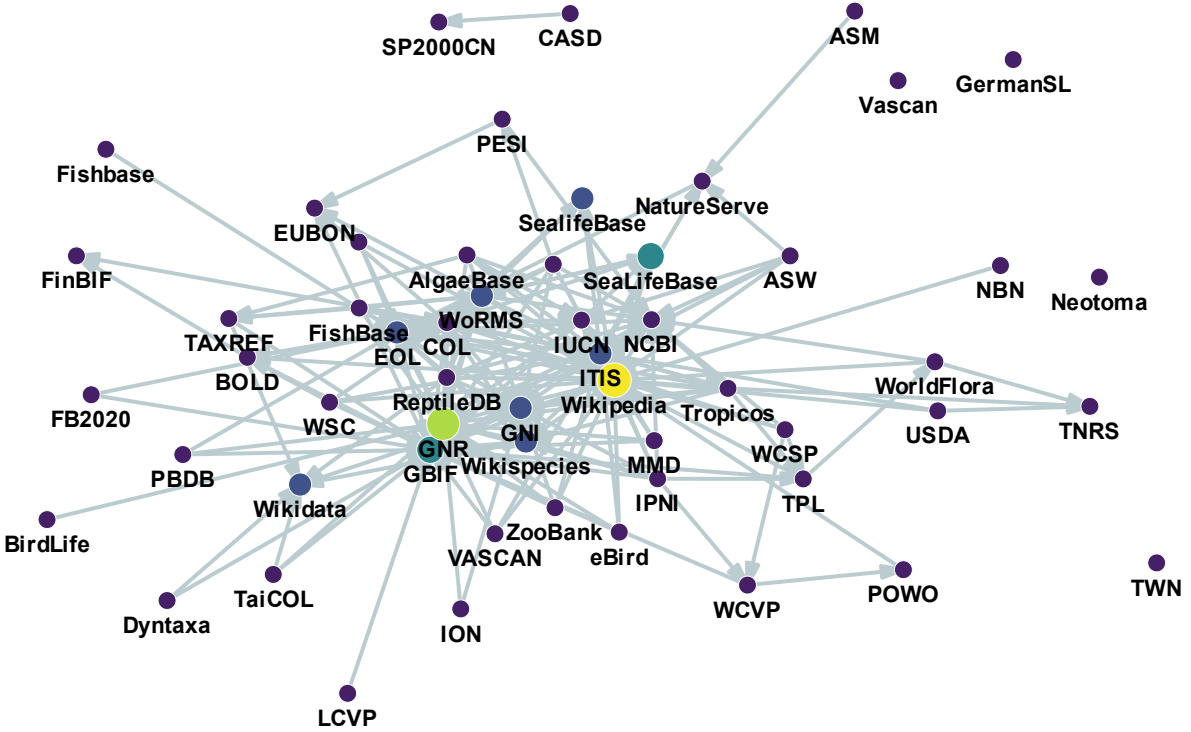


Beware of connections across databases

Relationships
Source → Target

Specific databases
widely reused

Some "mega-"
aggregator databases



N.B.: This information is **very difficult to gather**
(not even talking quantitatively)

Then we need the tools: R packages

Then we need the tools: R packages

Review of packages on CRAN + GitHub + Bioconductor

Then we need the tools: R packages

Review of packages on CRAN + GitHub + Bioconductor

Manually curated from standard queries (GitHub code search + r-pkg.org)

Then we need the tools: R packages

Review of packages on CRAN + GitHub + Bioconductor

Manually curated from standard queries (GitHub code search + r-pkg.org)

Criteria

Then we need the tools: R packages

Review of packages on CRAN + GitHub + Bioconductor

Manually curated from standard queries (GitHub code search + r-pkg.org)

Criteria

- **Real R package** (not collection of scripts)

Then we need the tools: R packages

Review of packages on CRAN + GitHub + Bioconductor

Manually curated from standard queries (GitHub code search + r-pkg.org)

Criteria

- **Real R package** (not collection of scripts)
- **Not a wrapper** (not only reusing other tools)

Then we need the tools: R packages

Review of packages on CRAN + GitHub + Bioconductor

Manually curated from standard queries (GitHub code search + r-pkg.org)

Criteria

- **Real R package** (not collection of scripts)
- **Not a wrapper** (not only reusing other tools)
- **Not about genomics** (different field)

Then we need the tools: R packages

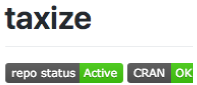
Review of packages on CRAN + GitHub + Bioconductor

Manually curated from standard queries (GitHub code search + r-pkg.org)

Criteria

- **Real R package** (not collection of scripts)
- **Not a wrapper** (not only reusing other tools)
- **Not about genomics** (different field)

Identified **68 packages** of which **59** were **included**



Packages for different purposes

Packages for different purposes

Identified **59 packages** belonging to **four categories**

Packages for different purposes

Identified **59 packages** belonging to **four categories**

Infrastructure

3 pkgs

taxa provides standard classes

Packages for different purposes

Identified **59 packages** belonging to **four categories**

Infrastructure

3 pkgs

`taxa` provides standard classes

Database Access

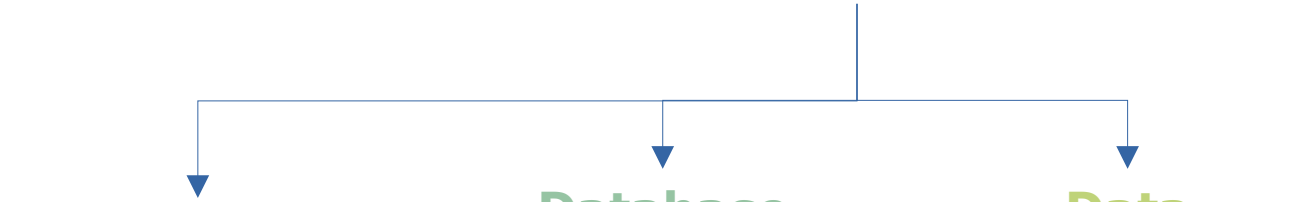
46 pkgs

`taxize` accesses many databases

`taxadb` offers offline access

Packages for different purposes

Identified **59 packages** belonging to **four categories**



Infrastructure

3 pkgs

`taxa` provides standard classes

Database Access

46 pkgs

`taxize` accesses many databases

`taxadb` offers offline access

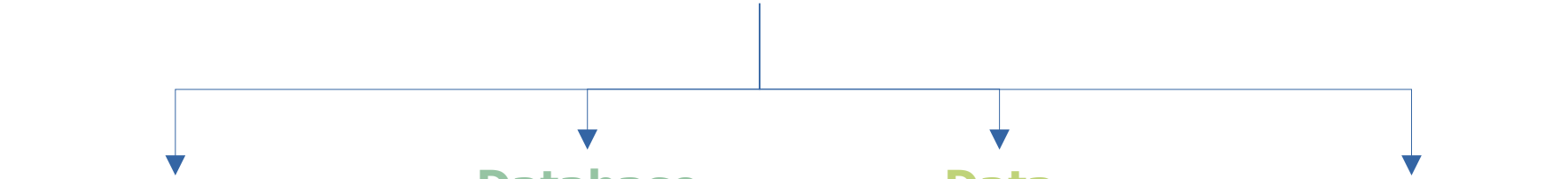
Data Wrangling

11 pkgs

`rgnparser` parses names into components

Packages for different purposes

Identified **59 packages** belonging to **four categories**



Infrastructure

3 pkgs

taxa provides standard classes

Database Access

46 pkgs

taxize accesses many databases

taxadb offers offline access

Data Wrangling

11 pkgs

rgnparser parses names into components

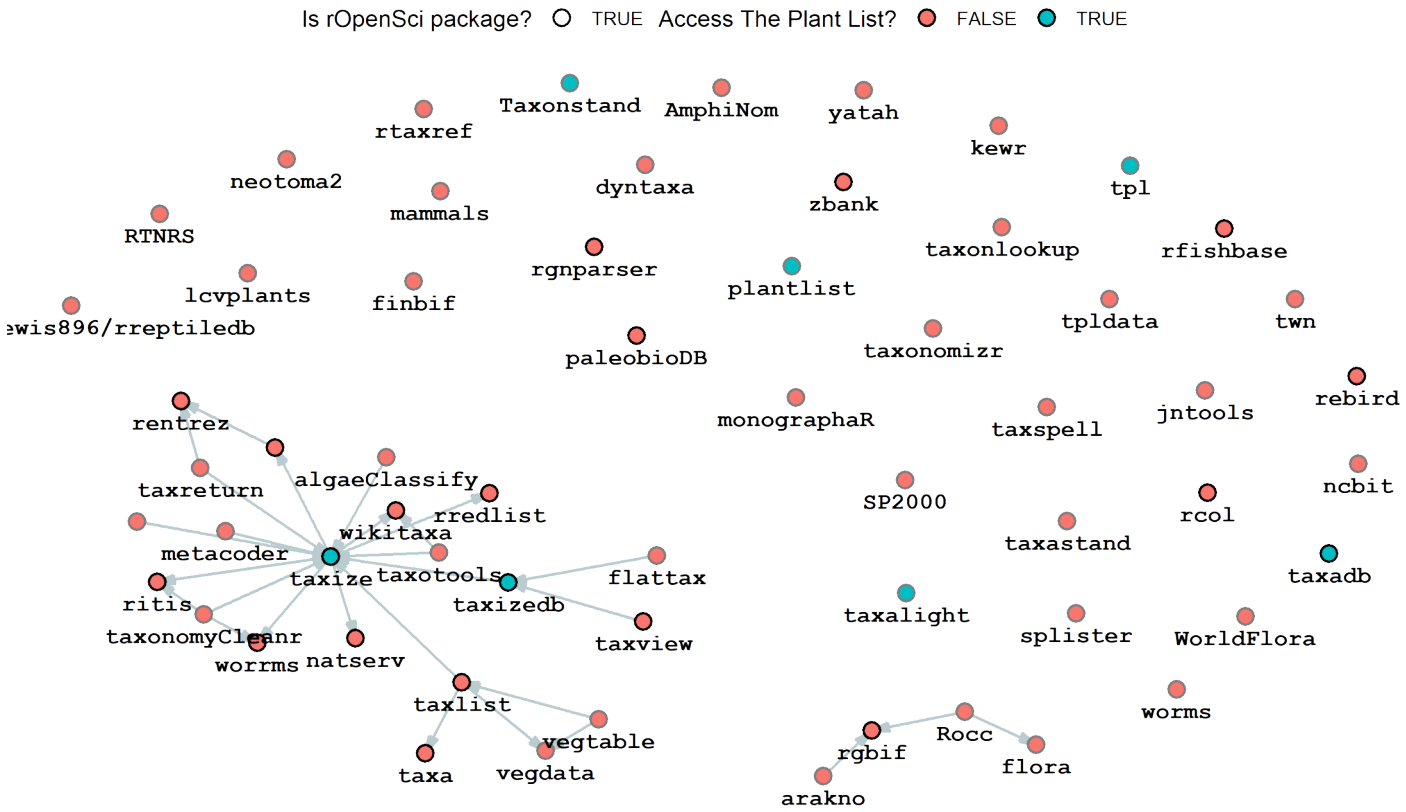
Visualization

2 pkgs

metacoder displays taxonomy

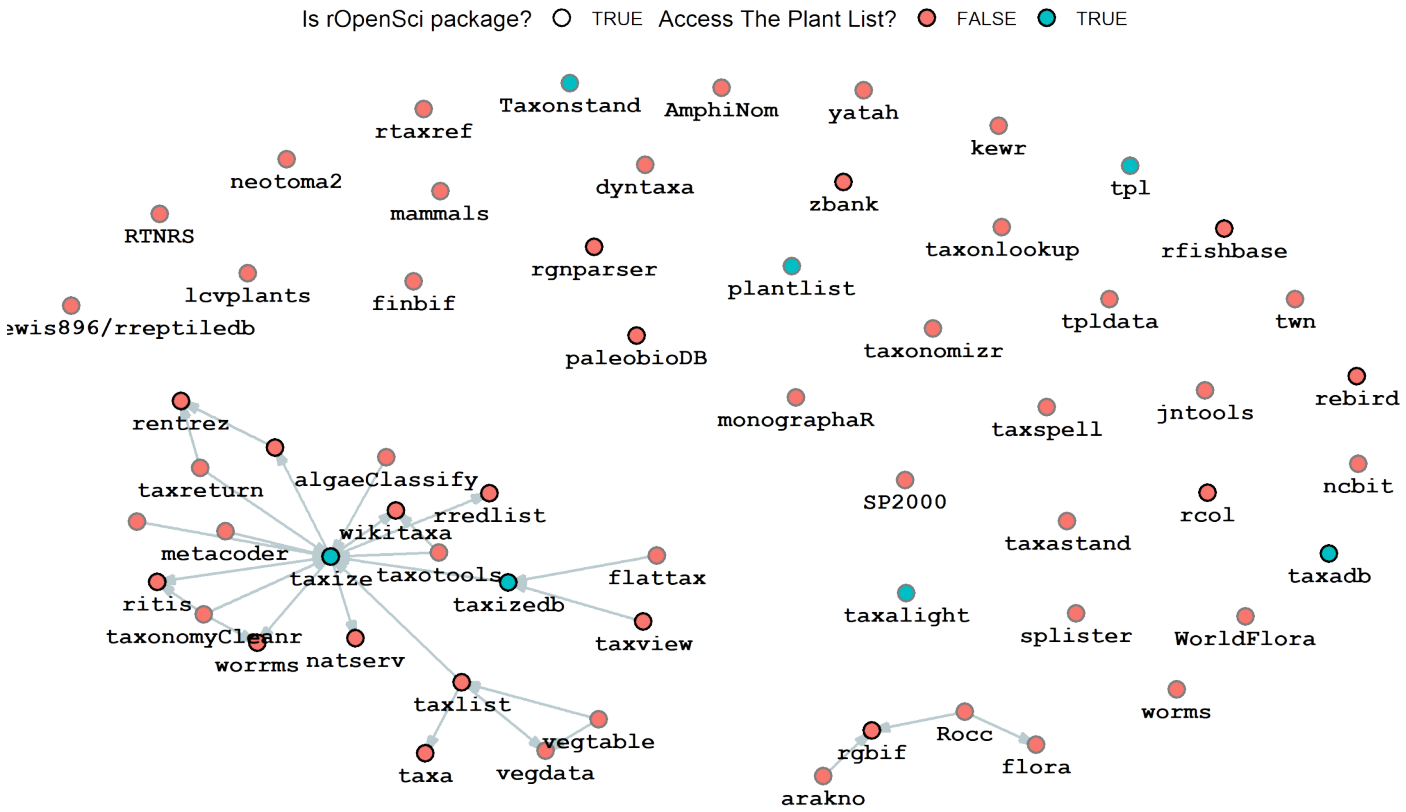
The Package Dependency Network

The Package Dependency Network



The Package Dependency Network

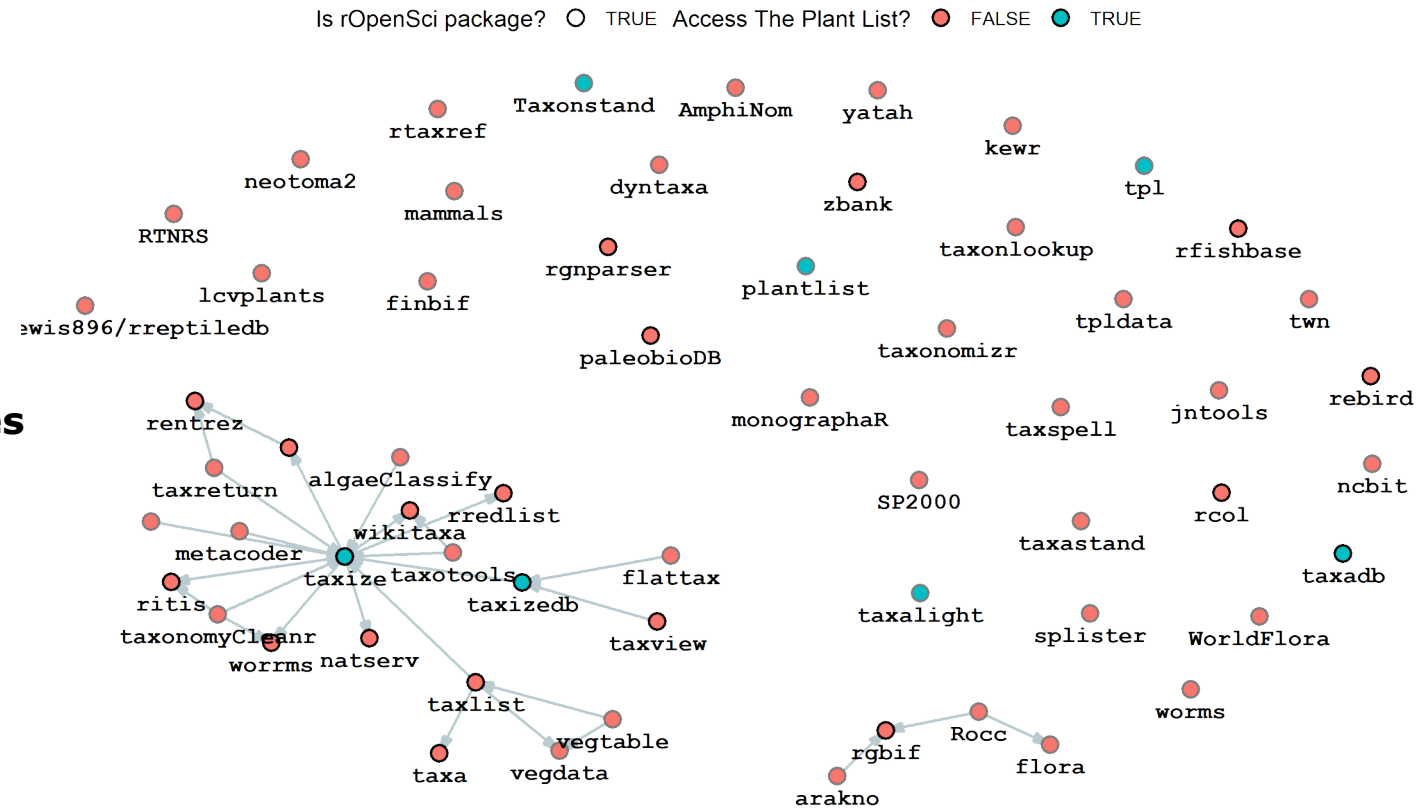
Disconnected network



The Package Dependency Network

Disconnected network

Apart from **rOpenSci packages**

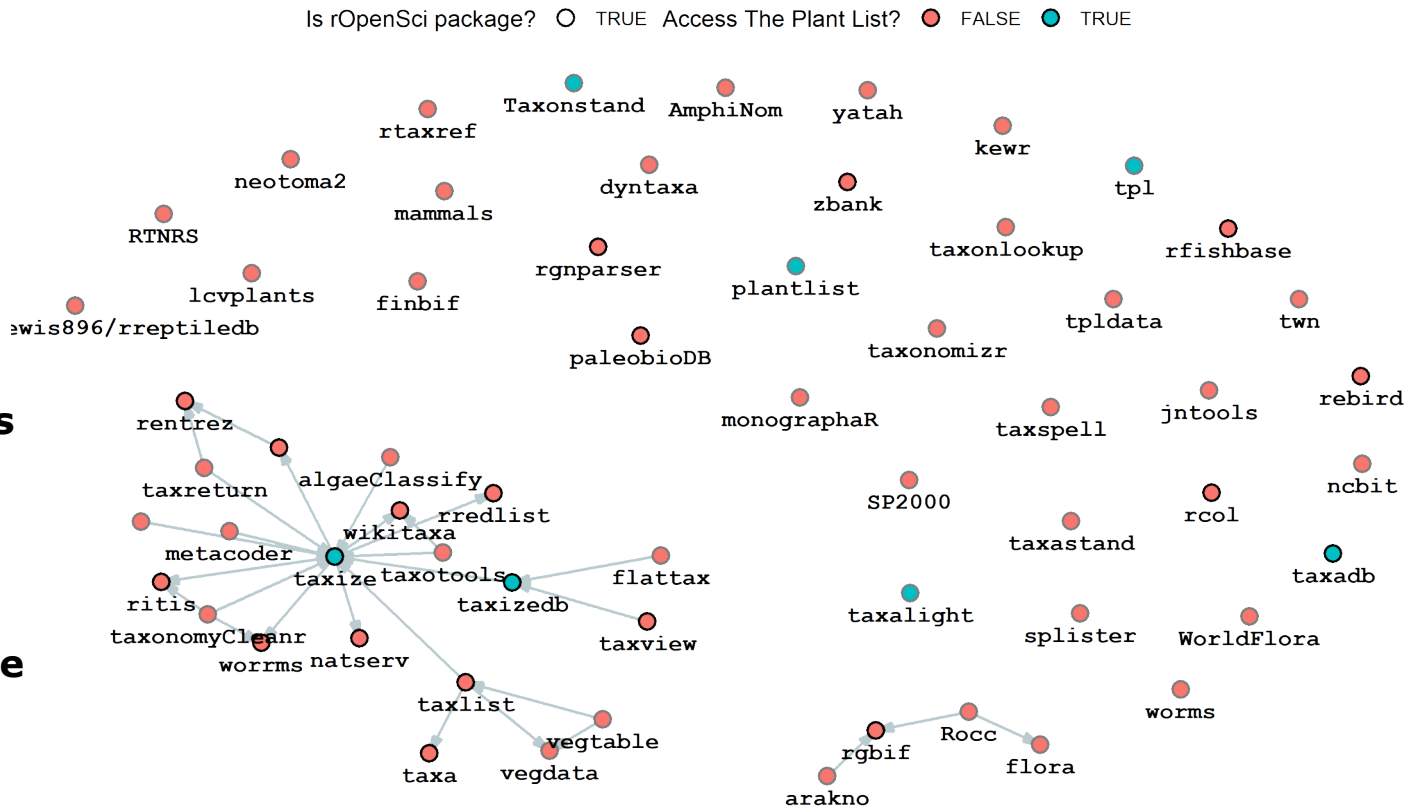


The Package Dependency Network

Disconnected network

Apart from **rOpenSci packages**

Many pkgs provide **same source** (TPL)



Taxharmonizexplorer: Connectining Sources & Packages

DEMO TIME!

https://mgrenie.shinyapps.io/taxharmonizexplorer/

Summary of this part

Summary of this part

- **Taxonomic Harmonization** requires **taxonomic reference databases**

Summary of this part

- **Taxonomic Harmonization** requires **taxonomic reference databases**
- These databases cover various **spatial scales** and **taxonomic breadth**

Summary of this part

- **Taxonomic Harmonization** requires **taxonomic reference databases**
- These databases cover various **spatial scales** and **taxonomic breadth**
- Some **larger databases** aggregate **smaller databases**

Summary of this part

- **Taxonomic Harmonization** requires **taxonomic reference databases**
- These databases cover various **spatial scales** and **taxonomic breadth**
- Some **larger databases** aggregate **smaller databases**
- **Taxonomic harmonization** also rely on tools which are diverse

Summary of this part

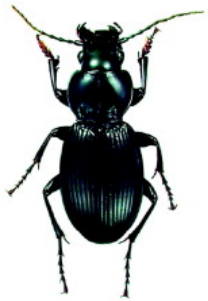
- **Taxonomic Harmonization** requires **taxonomic reference databases**
- These databases cover various **spatial scales** and **taxonomic breadth**
- Some **larger databases** aggregate **smaller databases**
- **Taxonomic harmonization** also rely on tools which are diverse
- Some **R packages** (e.g. taxize) **access many databases**

How to perform Taxonomic Harmonization?

QUESTION
Naively, how would you perform taxonomic harmonization?

Okay, I have species names, how do I proceed?

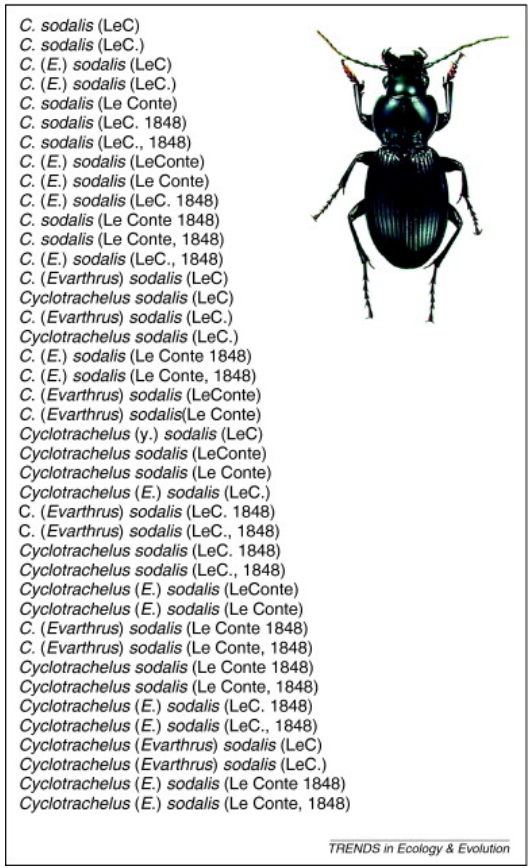
C. sodalis (LeC)
C. sodalis (LeC.)
C. (*E.*) *sodalis* (LeC)
C. (*E.*) *sodalis* (LeC.)
C. sodalis (Le Conte)
C. sodalis (LeC. 1848)
C. sodalis (LeC., 1848)
C. (*E.*) *sodalis* (LeConte)
C. (*E.*) *sodalis* (Le Conte)
C. (*E.*) *sodalis* (LeC. 1848)
C. sodalis (Le Conte 1848)
C. sodalis (Le Conte, 1848)
C. (*E.*) *sodalis* (LeC., 1848)
C. (*Evarthrus*) *sodalis* (LeC)
Cyclotrachelus sodalis (LeC)
C. (*Evarthrus*) *sodalis* (LeC.)
Cyclotrachelus sodalis (LeC.)
C. (*E.*) *sodalis* (Le Conte 1848)
C. (*E.*) *sodalis* (Le Conte, 1848)
C. (*Evarthrus*) *sodalis* (LeConte)
C. (*Evarthrus*) *sodalis*(Le Conte)
Cyclotrachelus (y.) *sodalis* (LeC)
Cyclotrachelus sodalis (LeConte)
Cyclotrachelus sodalis (Le Conte)
Cyclotrachelus (*E.*) *sodalis* (LeC.)
C. (*Evarthrus*) *sodalis* (LeC. 1848)
C. (*Evarthrus*) *sodalis* (LeC., 1848)
Cyclotrachelus sodalis (LeC. 1848)
Cyclotrachelus sodalis (LeC., 1848)
Cyclotrachelus (*E.*) *sodalis* (LeConte)
Cyclotrachelus (*E.*) *sodalis* (Le Conte)
C. (*Evarthrus*) *sodalis* (Le Conte 1848)
C. (*Evarthrus*) *sodalis* (Le Conte, 1848)
Cyclotrachelus sodalis (Le Conte 1848)
Cyclotrachelus sodalis (Le Conte, 1848)
Cyclotrachelus (*E.*) *sodalis* (LeC. 1848)
Cyclotrachelus (*E.*) *sodalis* (LeC., 1848)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeC)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeC.)
Cyclotrachelus (*E.*) *sodalis* (Le Conte 1848)
Cyclotrachelus (*E.*) *sodalis* (Le Conte, 1848)



TRENDS in Ecology & Evolution

Patterson et al. 2010 *TREE*

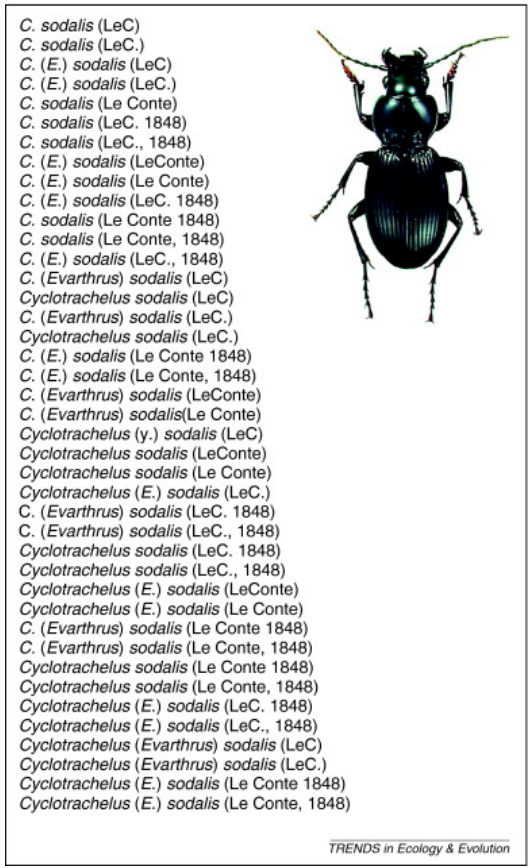
Okay, I have species names, how do I proceed?



Patterson et al. 2010 *TREE*

Diversity of writing **styles**

Okay, I have species names, how do I proceed?

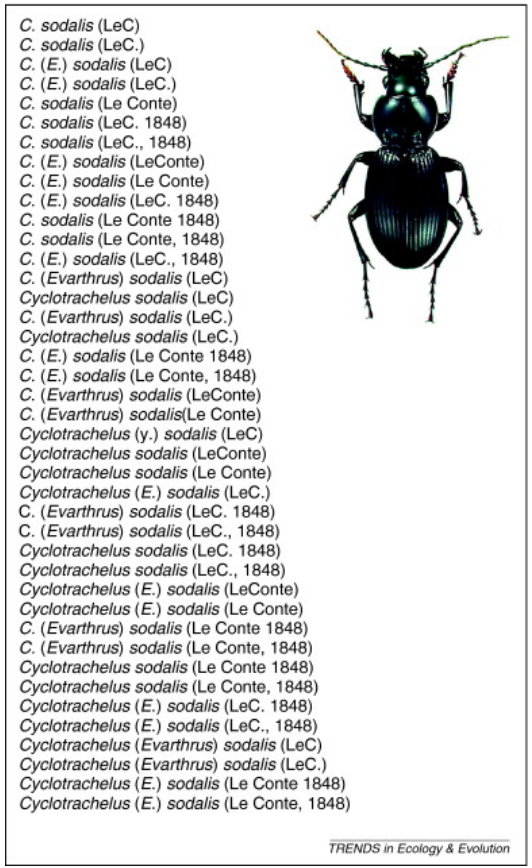


Patterson et al. 2010 *TREE*

Diversity of writing **styles**

Need to **standardize** it

Okay, I have species names, how do I proceed?



Patterson et al. 2010 *TREE*

Diversity of writing **styles**

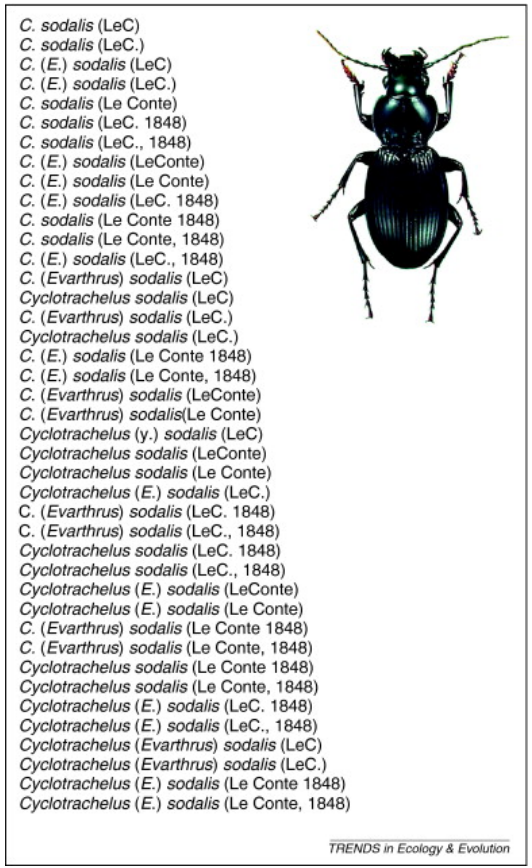
↓

Need to **standardize** it

↓

Identify **components** of names

Okay, I have species names, how do I proceed?



Patterson et al. 2010 *TREE*

Diversity of writing **styles**

↓

Need to **standardize** it

↓

Identify **components** of names

↓

Standardize **style**

Using Parsers to Identify Components of Names

Using Parsers to Identify Components of Names

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

Using Parsers to Identify Components of Names

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

44 326 unique taxa
in BioTIME

Using Parsers to Identify Components of Names

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

44 326 unique taxa in BioTIME → **32 900** after **pre-processing**

Using Parsers to Identify Components of Names

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

44 326 unique taxa in BioTIME → **32 900** after **pre-processing**

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

Using Parsers to Identify Components of Names

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

44 326 unique taxa in BioTIME → **32 900** after **pre-processing**

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

Also online:

Using Parsers to Identify Components of Names

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

44 326 unique taxa in BioTIME → **32 900** after **pre-processing**

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

Also online:
https://parser.globalnames.org/
https://www.gbif.org/tools/name-parser

No one-size-fit-all BUT tested four workflows

No one-size-fit-all BUT tested four workflows

1: Pre-process
(unify spelling)

No one-size-fit-all BUT tested four workflows

1: Pre-process
(unify spelling)

2: Match databases

No one-size-fit-all BUT tested four workflows

1: Pre-process
(unify spelling)

2: Match databases

3: Harmonize

No one-size-fit-all BUT tested four workflows

1: Pre-process
(unify spelling)

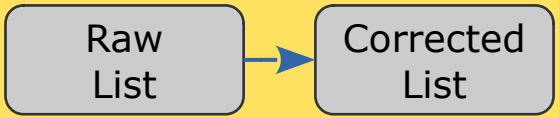
Raw
List

2: Match databases

3: Harmonize

No one-size-fit-all BUT tested four workflows

1: Pre-process
(unify spelling)



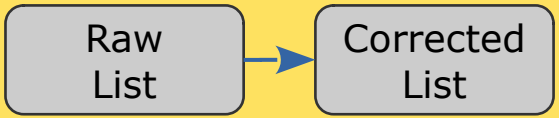
2: Match databases

3: Harmonize

No one-size-fit-all BUT tested four workflows

1: Pre-process
(unify spelling)

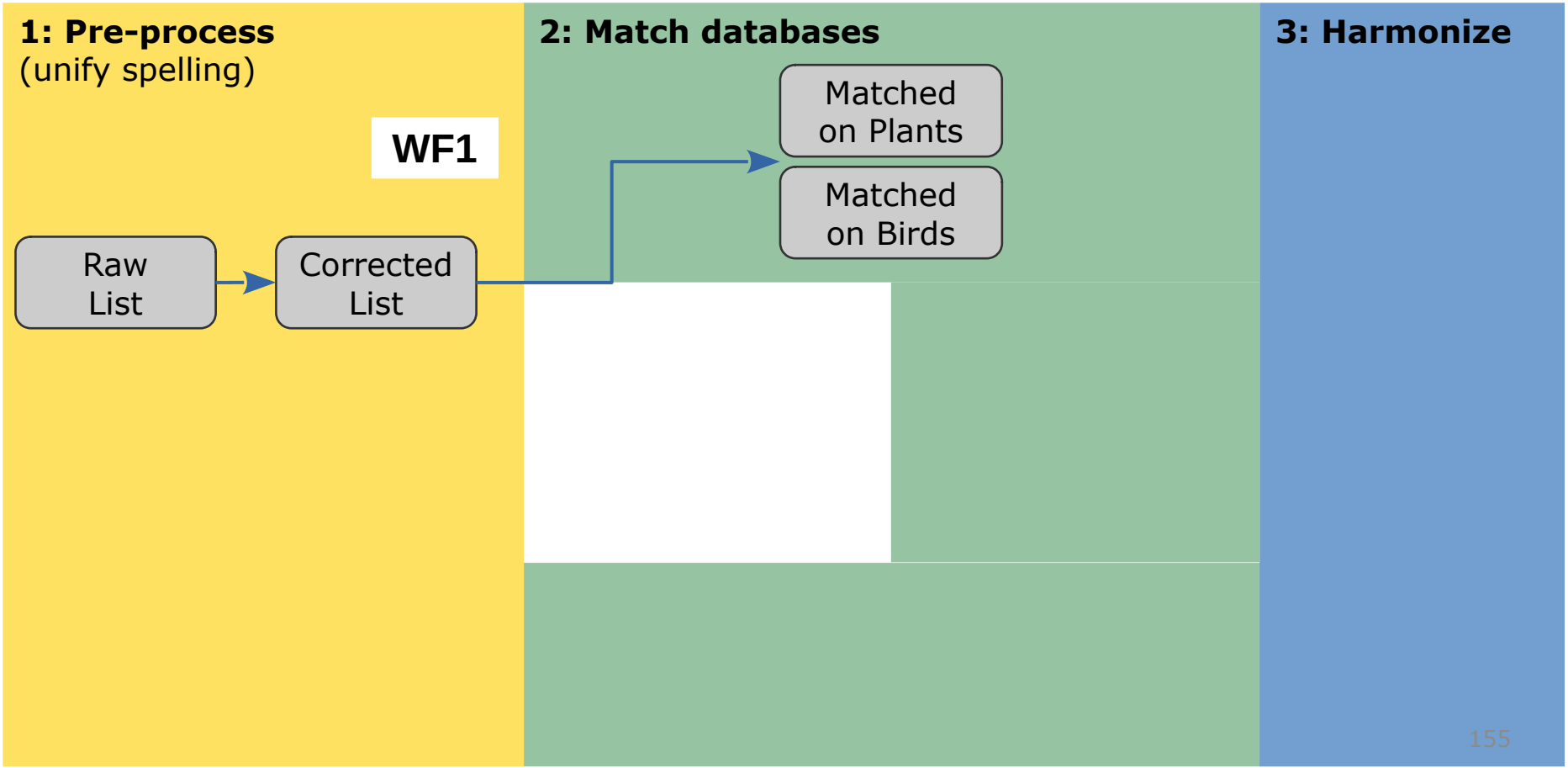
WF1



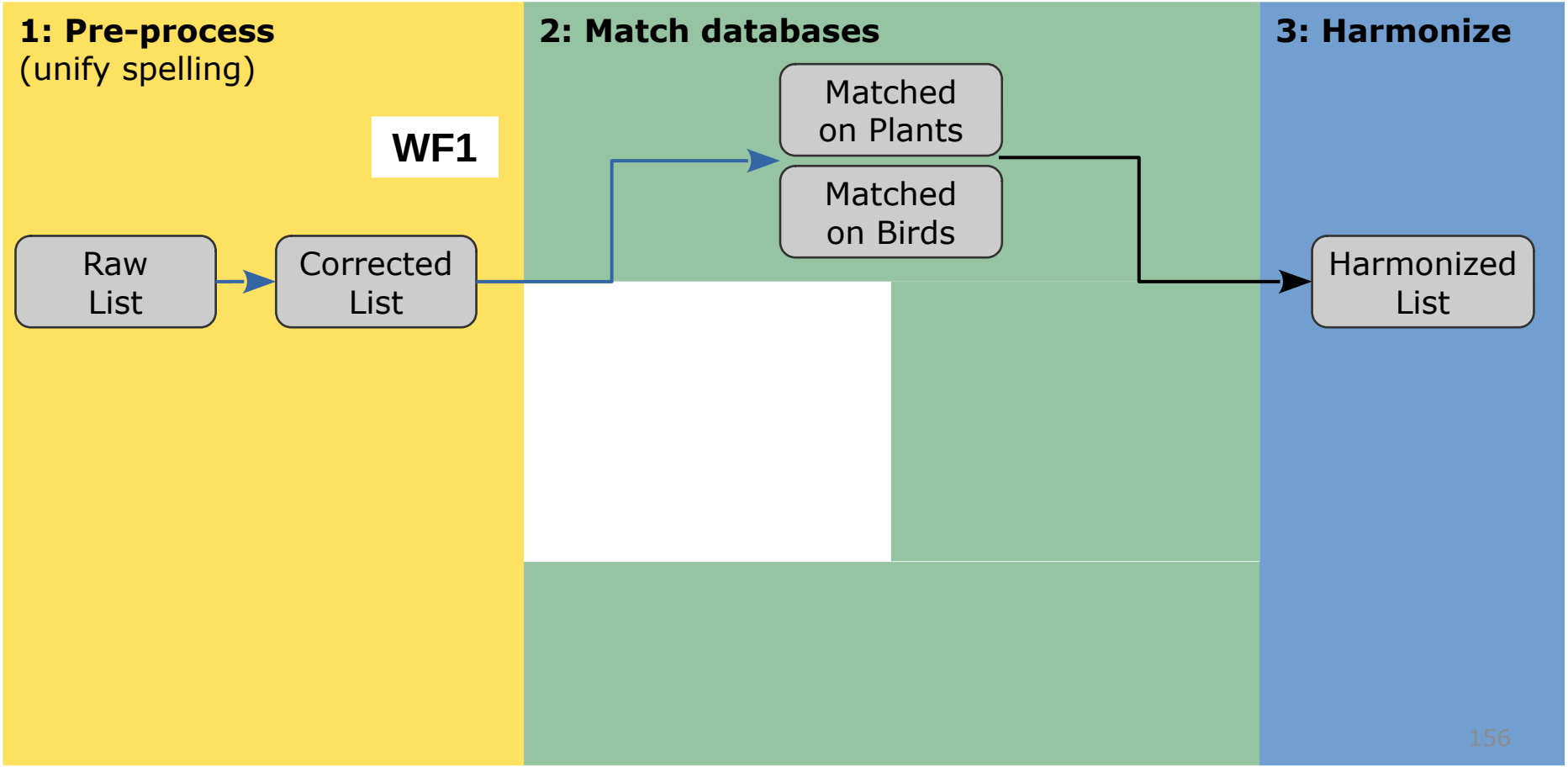
2: Match databases

3: Harmonize

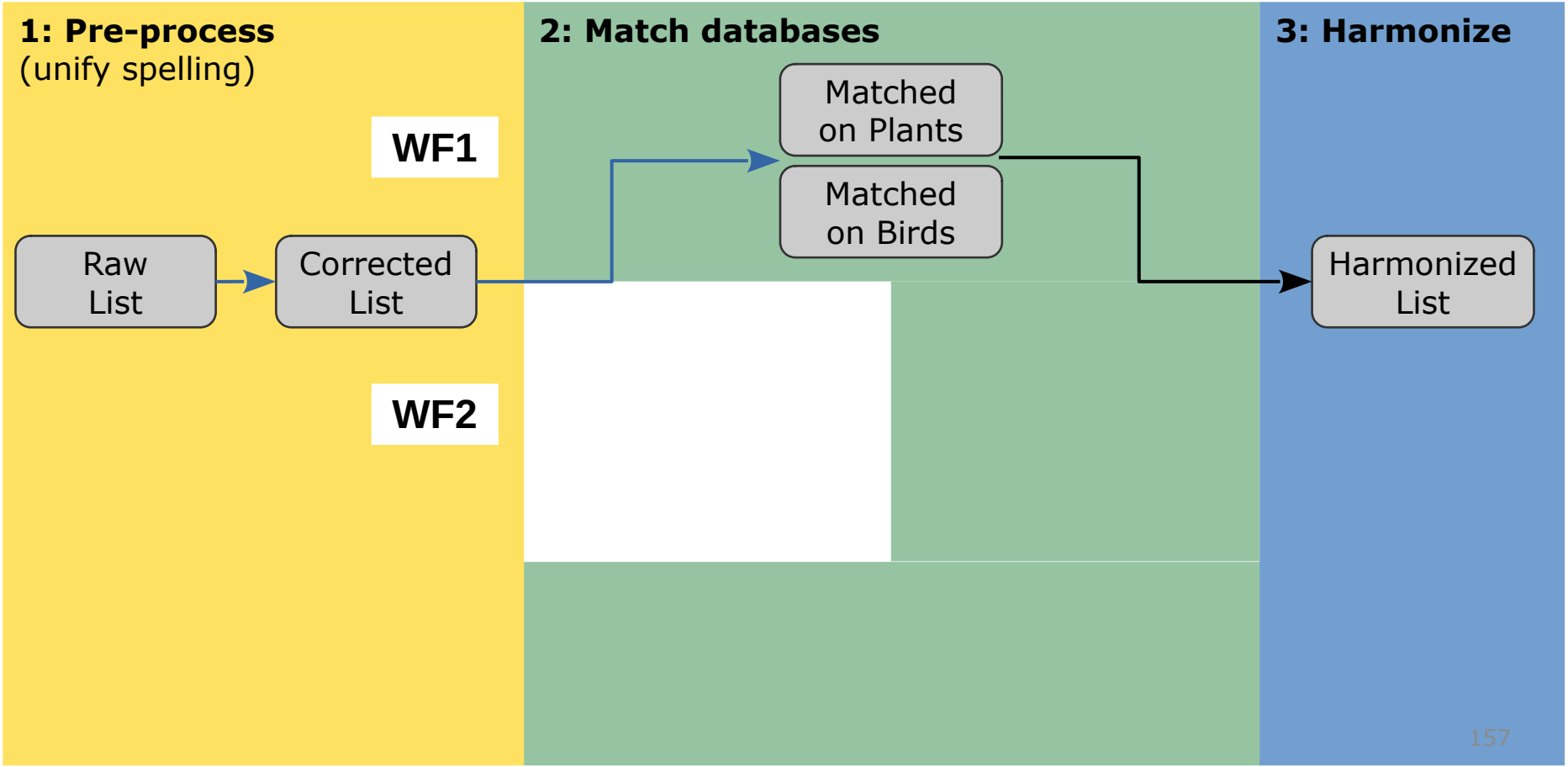
No one-size-fit-all BUT tested four workflows



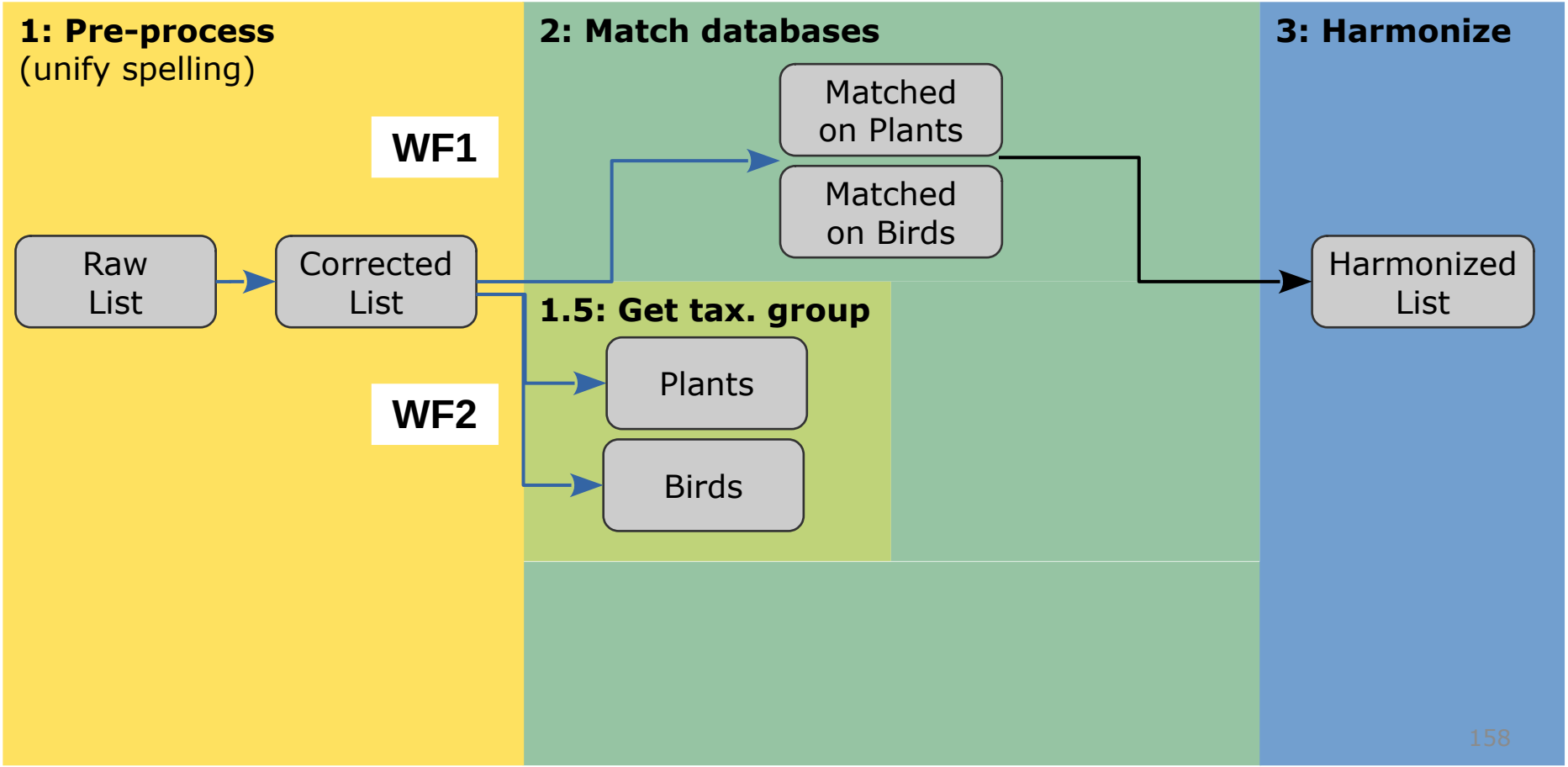
No one-size-fit-all BUT tested four workflows



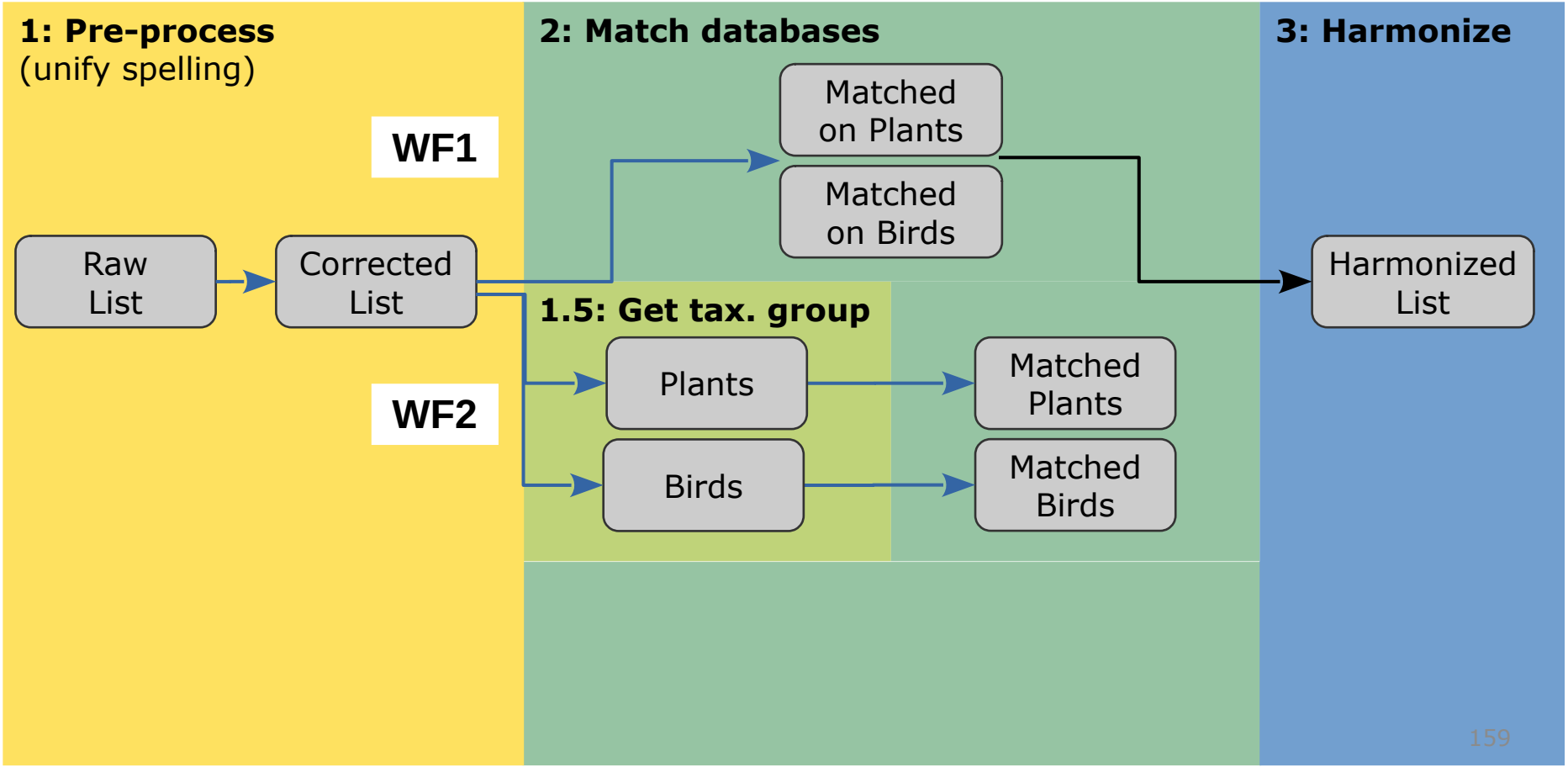
No one-size-fit-all BUT tested four workflows



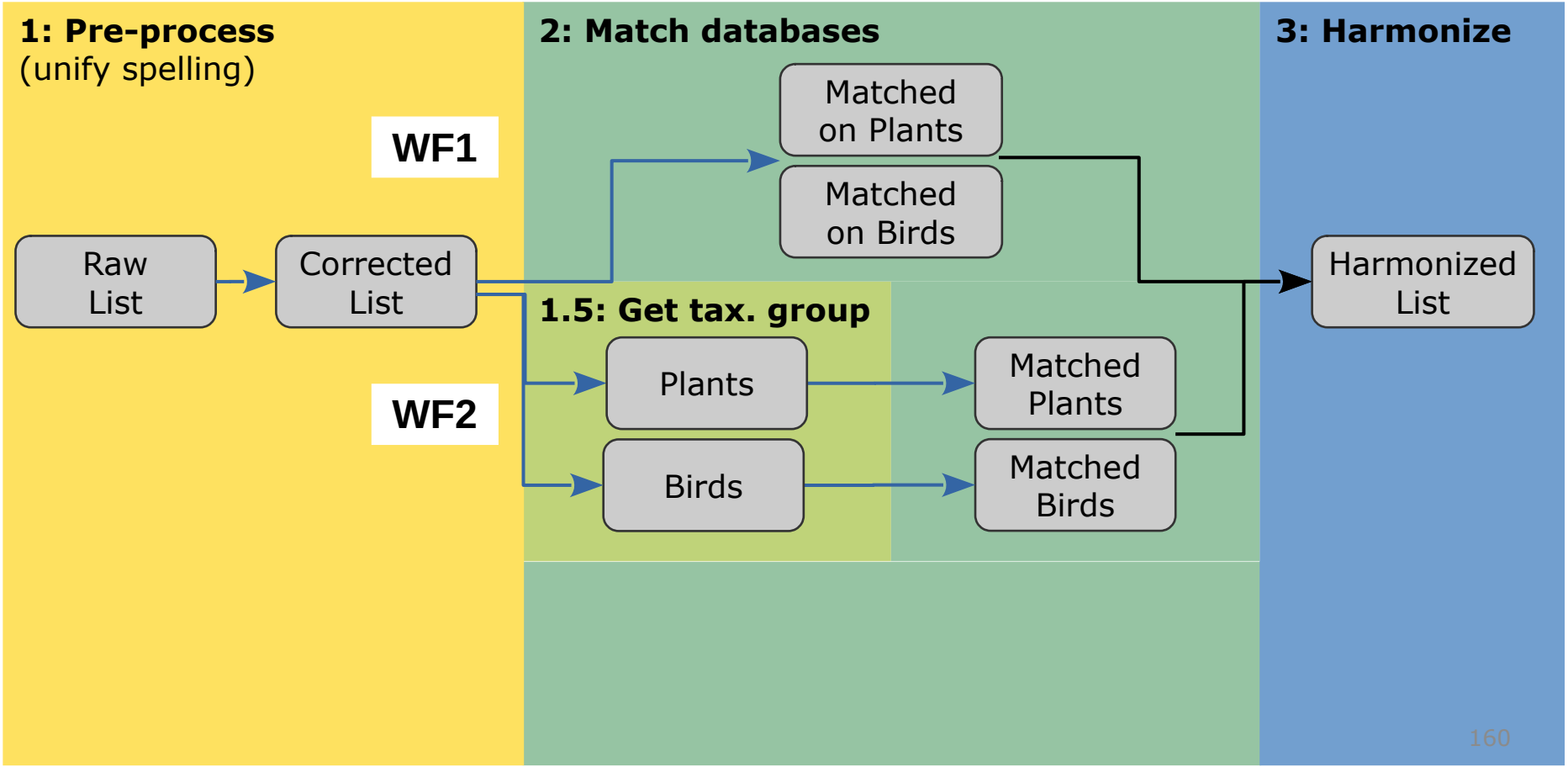
No one-size-fit-all BUT tested four workflows



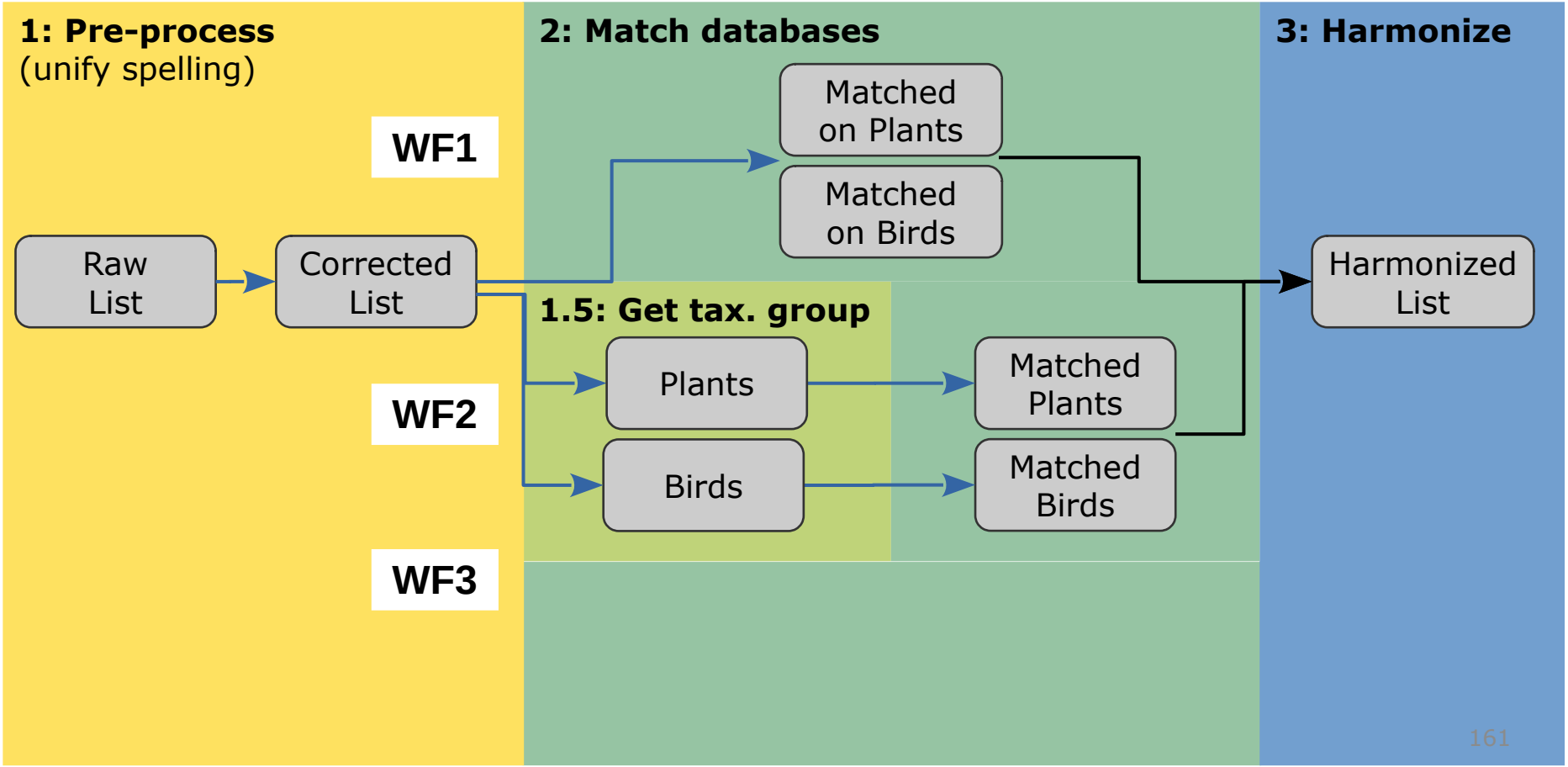
No one-size-fit-all BUT tested four workflows



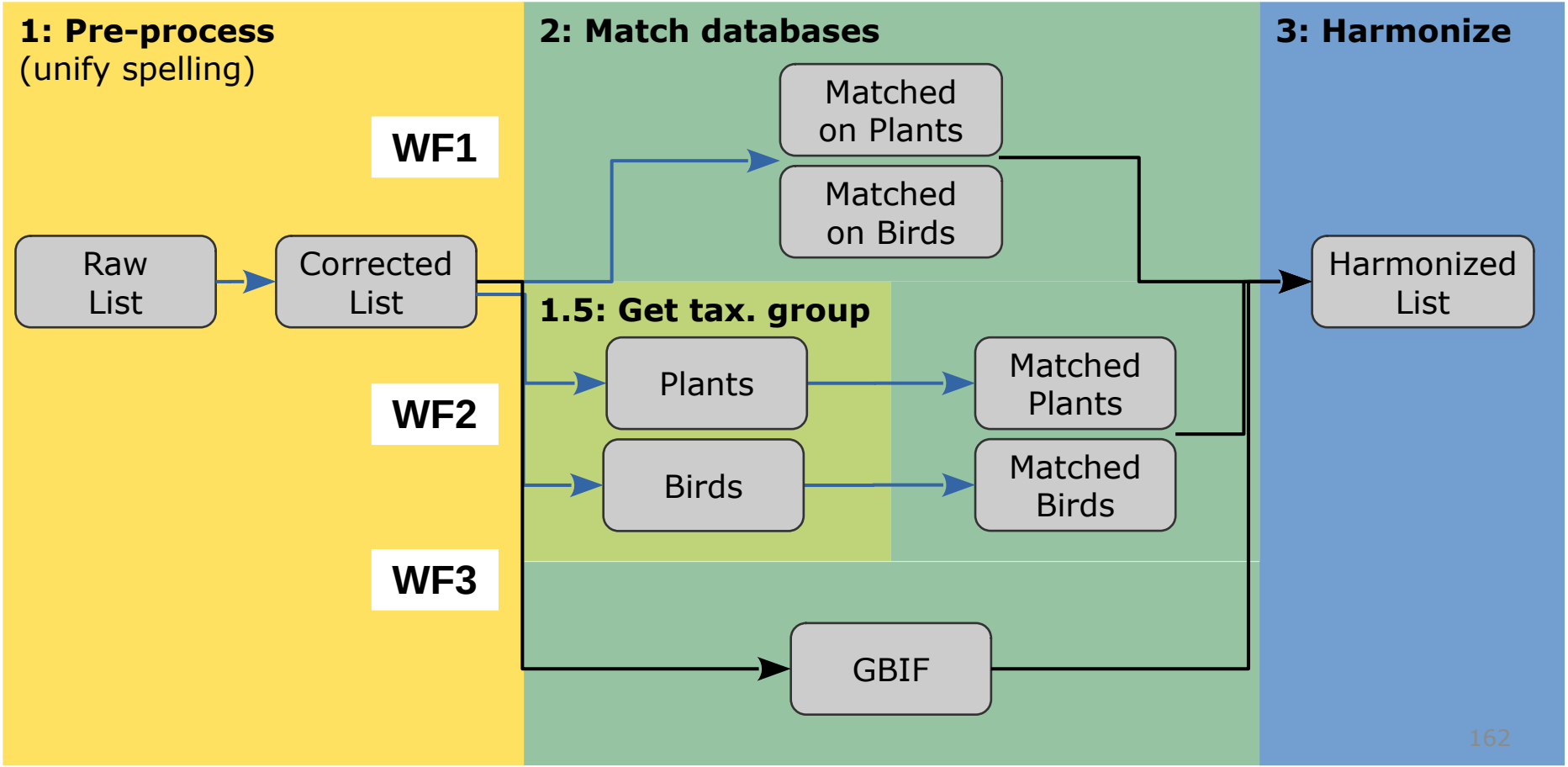
No one-size-fit-all BUT tested four workflows



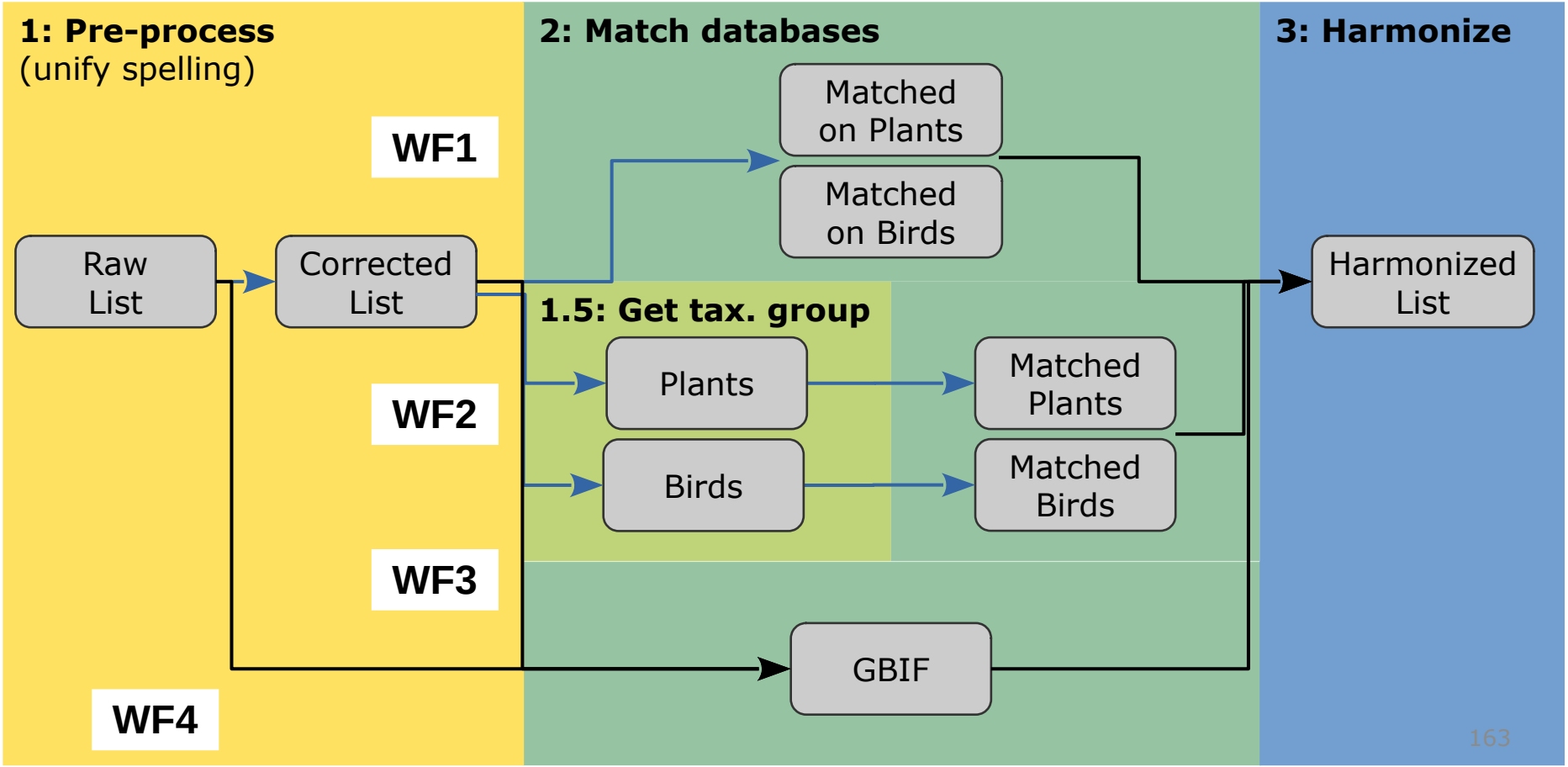
No one-size-fit-all BUT tested four workflows



No one-size-fit-all BUT tested four workflows



No one-size-fit-all BUT tested four workflows



1. Check encoding and special characters

2. Parse name into components

3. Match names against well-selected databases

3. Match names against well-selected databases

Select databases based on:

3. Match names against well-selected databases

Select databases based on:

1. Taxonomic breadth (more specialized have more details)

3. Match names against well-selected databases

Select databases based on:

1. Taxonomic breadth (more specialized have more details)

2. Spatial Scale (smaller databases tend to be more up-to-date)

3. Match names against well-selected databases

Select databases based on:

1. Taxonomic breadth (more specialized have more details)

2. Spatial Scale (smaller databases tend to be more up-to-date)

3. Date of update (more recent better reflect current knowledge)

3. Match names against well-selected databases

Select databases based on:

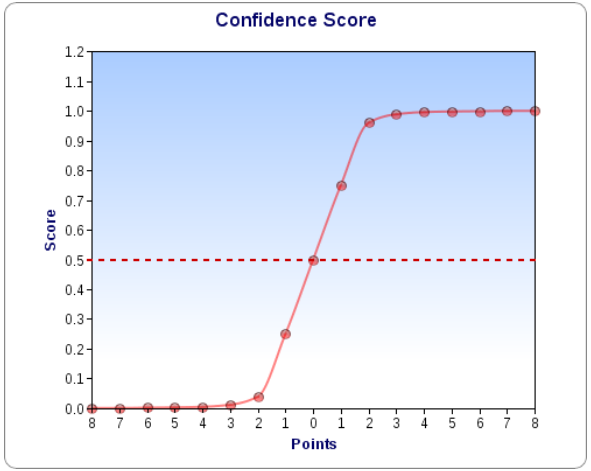
1. Taxonomic breadth (more specialized have more details)
2. Spatial Scale (smaller databases tend to be more up-to-date)
3. Date of update (more recent better reflect current knowledge)
4. Number of synonyms (more synonyms means better resolving)

4. Resolve them

Rubus rubus

Rubus rubus [exact match, Score: 0.988]

CU*STAR

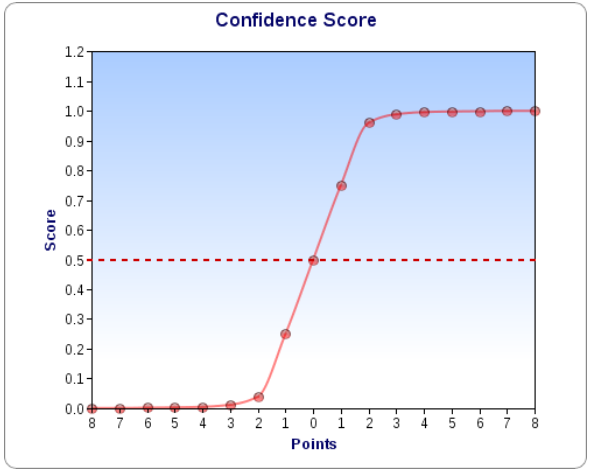


4. Resolve them

Rubus rubus

Rubus rubus [exact match, Score: 0.988]

CU*STAR



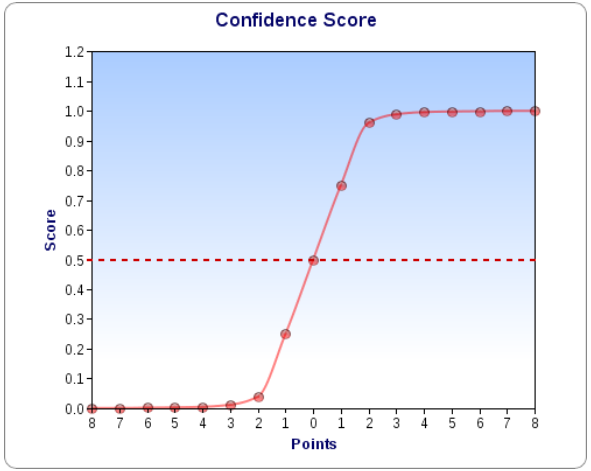
Use **matching scores** to resolve matching names automatically

4. Resolve them

Rubus rubus

Rubus rubus [exact match, Score: 0.988]

CU*STAR



Use **matching scores** to resolve matching names automatically

Not perfect but difficult to achieve with many (>1,000) names

Summary of the workflow

Summary of the workflow

1. Check encoding and special characters

Summary of the workflow

1. Check encoding and special characters
2. Parse name into components

Summary of the workflow

1. Check encoding and special characters
2. Parse name into components
3. Match names against well-selected databases

Summary of the workflow

1. Check encoding and special characters
2. Parse name into components
3. Match names against well-selected databases
4. Resolve names

What to do with unmatched names?

What to do with unmatched names?

1. Check spelling manually

What to do with unmatched names?

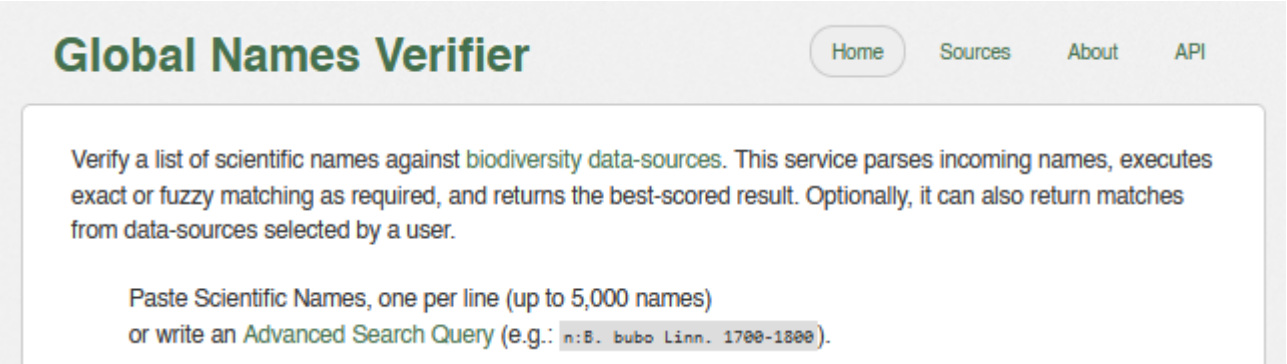
1. Check spelling manually

2. Drop them?

What to do with unmatched names?

1. Check spelling manually
2. Drop them?
3. Test against “aggregator databases” like Global Names Verifier

https://verifier.globalnames.org/



Other points of detail: author names

Other points of detail: author names

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

Other points of detail: author names

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

Gives scientific information

Other points of detail: author names

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

Gives scientific information

Help **distinguish** different **species concepts**

Other points of detail: author names

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

Gives scientific information

Help **distinguish** different
species concepts

→ Whenever possible (not all databases allow it)
include author names

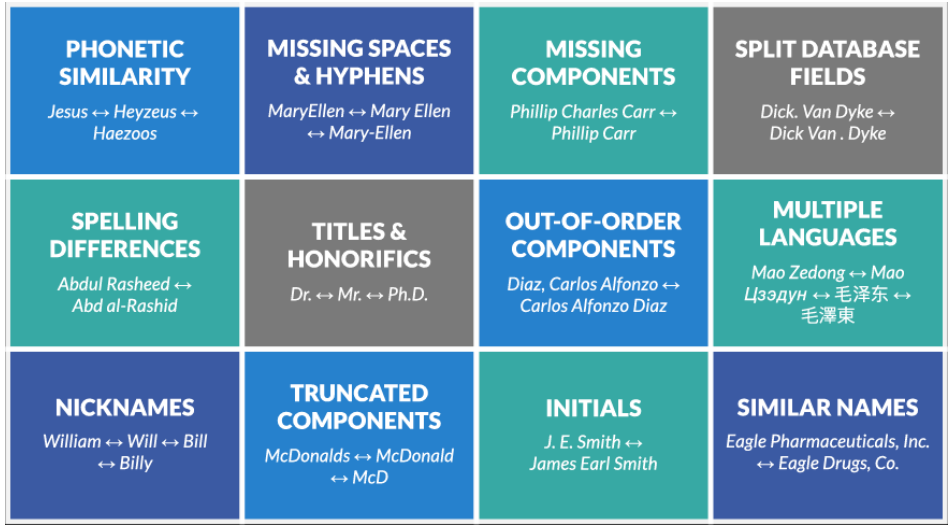
Other points of detail: fuzzy matching

Other points of detail: fuzzy matching

Fuzzy matching means **matching** names that are **not exactly written the same way**

Other points of detail: fuzzy matching

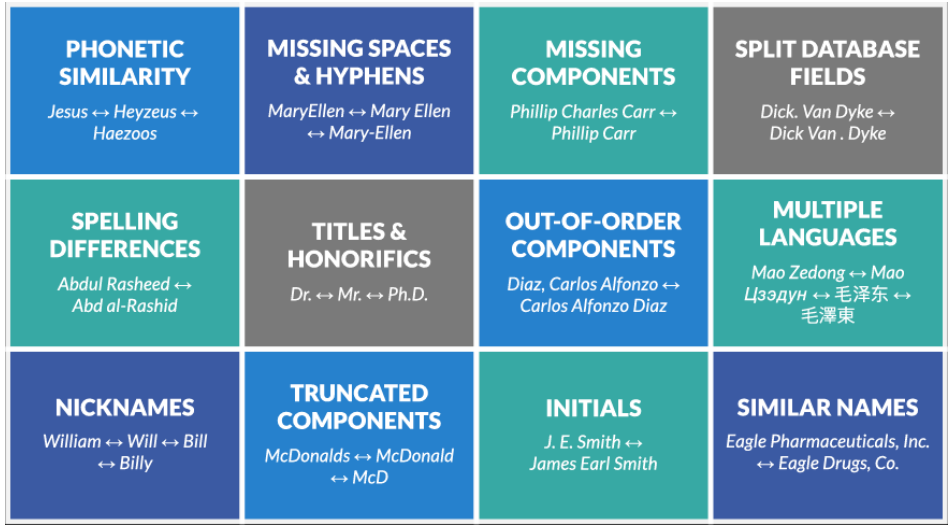
Fuzzy matching means **matching** names that are **not exactly written the same way**



Variety of techniques implemented by databases

Other points of detail: fuzzy matching

Fuzzy matching means **matching** names that are
not exactly written the same way



Variety of techniques
implemented by databases

→ Can use it but with care (can match outside taxa or being too loose)

Summary of this part

Summary of this part

- Taxonomic Harmonization has **no one-size-fit-all solution**

Summary of this part

- Taxonomic Harmonization has **no one-size-fit-all solution**
- Checking **data encoding** and **parsing name components** is important

Summary of this part

- Taxonomic Harmonization has **no one-size-fit-all solution**
- Checking **data encoding** and **parsing name components** is important
- Then **match** names on **appropriate databases**

Summary of this part

- Taxonomic Harmonization has **no one-size-fit-all solution**
- Checking **data encoding** and **parsing name components** is important
- Then **match** names on **appropriate databases**
- **Resolve** them based on **matching scores**

Summary of this part

- Taxonomic Harmonization has **no one-size-fit-all solution**
- Checking **data encoding** and **parsing name components** is important
- Then **match** names on **appropriate databases**
- **Resolve** them based on **matching scores**
- **Decide** what to do about **unmatched names**

Take Home Messages and Grand Summary of Things

The Grand Summary of Things

The Grand Summary of Things

- Taxonomic harmonization is the process of matching taxonomic names onto reference databases

The Grand Summary of Things

- Taxonomic harmonization is the process of matching taxonomic names onto reference databases
- It’s paramount for the data matching process when combining data

The Grand Summary of Things

- Taxonomic harmonization is the process of matching taxonomic names onto reference databases
- It’s paramount for the data matching process when combining data
- It is supported by databases of different spatial scale and taxonomic breadth, as well as a diversity of tools

The Grand Summary of Things

- Taxonomic harmonization is the process of matching taxonomic names onto reference databases
- It’s paramount for the data matching process when combining data
- It is supported by databases of different spatial scale and taxonomic breadth, as well as a diversity of tools
- Pre-processing the names is important for harmonizing taxonomy

The Grand Summary of Things

- Taxonomic harmonization is the process of matching taxonomic names onto reference databases
- It’s paramount for the data matching process when combining data
- It is supported by databases of different spatial scale and taxonomic breadth, as well as a diversity of tools
- Pre-processing the names is important for harmonizing taxonomy
- Think and decide about the matching process (scores, fuzzy matching, author names, etc.)

Remaining questions?

sDiv | synthesis centre of iDiv

Email: matthias.grenie@idiv.de
Twitter: @LeNematode
Mastodon: @LeNematode@pouet.chapril.org
Website: https://rekyt.github.io

Thank you!

NFDI 4 BIODIVERSITY
BIODIVERSITY, ECOLOGY & ENVIRONMENTAL DATA

GfÖ The Ecological Society of Germany, Austria and Switzerland