

# Data management - Data structuring



# Tips for data structuring



# McGill's 10 commandments

*Van Klink*

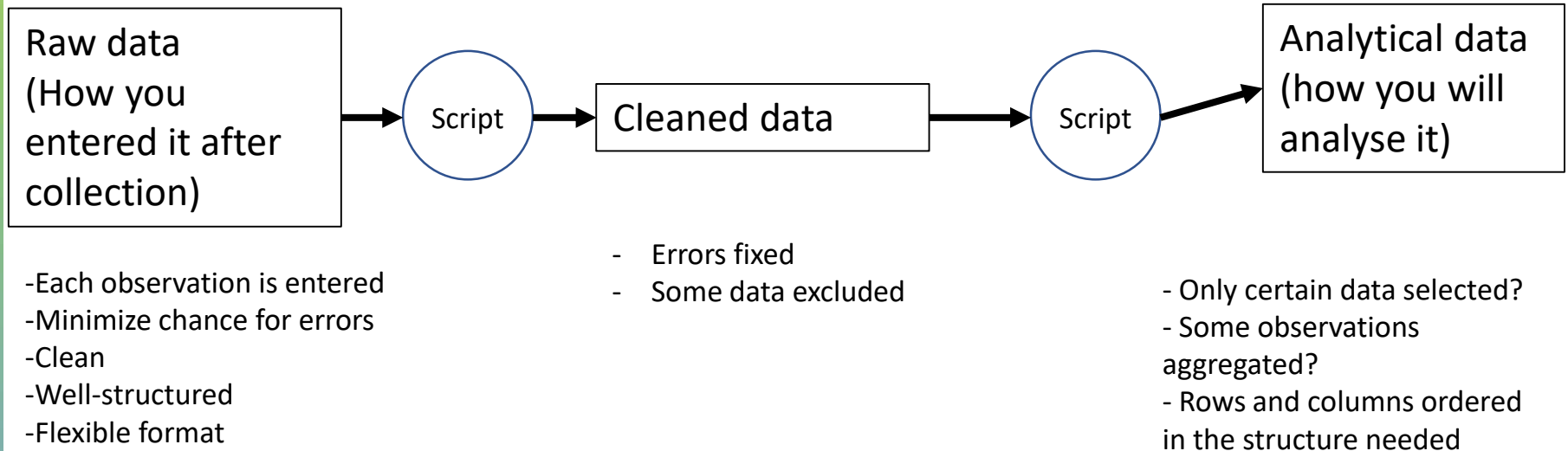
- 1 – Thou shalt distinguish raw data from analytical data and link them with a repeatable pipeline
- 5 – Thou shalt not hand-edit raw data
- 2 – Thou shalt create raw data in instance-row, variable-column format
- 3 -Thou shalt partially (but no more and no less) normalize raw data into star schema(s)
- 9 – Though shalt obsessively hand check the transformation from raw data to analytical data



1 – Thou shalt distinguish raw data from analytical data and link them with a repeatable pipeline

5 – Thou shalt not hand-edit raw data

*Van Klink*



## 2 – Thou shalt create raw data in instance-row, variable-column format

*Van Klink*



## 2 – Thou shalt create raw data in instance-row, variable-column format

*Van Klink*

- Clean and neat
- Each observation you make gets its own row (thus no data ever gets lost)
- Units of columns are clear
- Easy to add columns with notes or quality information
- Easy to convert to other formats (e.g. matrix) and to aggregate over several samples
- BUT:
  - More work during data entry
  - True or false absences are hard to distinguish
  - May not be the format you want for analysis (but this is easily solved)

|            | 2010 | 2011 | 2012 |
|------------|------|------|------|
| Washington | 20   | 22   | 18   |
| Oregon     | 10   | 13   | 5    |

| Site       | Year | Abundance |
|------------|------|-----------|
| Washington | 2010 | 20        |
| Washington | 2011 | 22        |
| Washington | 2012 | 18        |
| Oregon     | 2010 | 10        |
| Oregon     | 2011 | 13        |
| Oregon     | 2012 | 5         |



## 2 – Thou shalt create raw data in instance-row, variable-column format

Van Klink

Not like this:

| Sample | Acarina | Aeshnidae | Amphiagrion_sp. | Amphipoda | Ampullaridae | Ancylidae | Anisoptera | Anomalopsychidae | Aphylla |
|--------|---------|-----------|-----------------|-----------|--------------|-----------|------------|------------------|---------|
| a      | 0       | 0         | 0               | 0         | 0            | 1         | 0          | 2                | 0       |
| b      | 3       | 0         | 1               | 2         | 0            | 0         | 1          | 0                | 0       |
| c      | 0       | 0         | 0               | 3         | 0            | 0         | 0          | 1                | 45      |
| d      | 6       | 0         | 0               | 4         | 0            | 0         | 25         | 0                | 0       |
| e      | 1       | 0         | 0               | 5         | 3            | 0         | 40         | 0                | 0       |
| f      | 0       | 0         | 3               | 6         | 0            | 0         | 456        | 0                | 0       |
| g      | 9       | 0         | 0               | 0         | 0            | 0         | 0          | 1                | 0       |
| h      | 0       | 3         | 0               | 0         | 0            | 0         | 27         | 0                | 4       |
| i      | 3       | 0         | 0               | 0         | 0            | 0         | 0          | 0                | 1       |
| j      | 0       | 0         | 0               | 0         | 0            | 0         | 1          | 0                | 0       |

But like this:

| sample | Taxon           | Number |
|--------|-----------------|--------|
| a      | Ancylidae       | 1      |
| a      | Anomalopsychida | 2      |
| b      | Acarina         | 3      |
| b      | Amphiagrion_sp. | 1      |
| b      | Amphipoda       | 2      |
| b      | Anisoptera      | 1      |
| c      | Amphipoda       | 3      |
| c      | Anomalopsychida | 1      |
| c      | Aphylla         | 45     |
| ...    |                 |        |



# More commandments from Roel

Van Klink

- NEVER EVER put more than one piece of information in a cell, this is incredibly hard to separate later. The worst is when the pieces of information have inconsistent number of characters or no delimiters Example:

| Site                        | species           | Number |
|-----------------------------|-------------------|--------|
| Site5Plot28Makarere20150203 | PasserDomesticusM | 50     |
| Site5Plot28Makarere20150203 | PasserDomesticusF | 25     |
| Site5Plot28Makarere20150203 | TurdusMerulaF     | 1      |
| Site5Plot28Makarere20150203 | FalcoTinnunculusM | 1      |

- Never use multiple column headers example:

| Site 1 | Site 2 | Site 1 | Site 2 |
|--------|--------|--------|--------|
| Date 1 | Date 1 | Date 2 | Date 2 |
| ...    | ...    | ...    | ...    |
| ...    | ...    | ...    | ...    |

- Don't repeat information that can better be calculated e.g.: males, females, total

| spX_males | spX_females | Spx_total |
|-----------|-------------|-----------|
| ...       | ...         | ...       |





# 3 -Thou shalt partially (but no more and no less) normalize raw data into star schema(s)

Van Klink

## – Normalization

| Family   | Diet  | SpeciesName  | Abundance | IndivID | Weight_g | Sex |
|----------|-------|--------------|-----------|---------|----------|-----|
| Anatidae | Plant | Wood Duck    | 508.2     | LJ001   | 650      | M   |
| Anatidae | Plant | Mallard      | 5747.2    | LJ002   | 1050     | M   |
| Gaviidae | Fish  | Pacific Loon | 10        | LJ003   | 4126     | M   |
| Gaviidae | Fish  | Common Loon  | 277.4     | LJ004   | 4002     | F   |
| Anatidae | Plant | Wood Duck    | 508.2     | PD001   | 605      | F   |
| Anatidae | Plant | Mallard      | 5747.2    | PD004   | 1098     | M   |
| Anatidae | Plant | Wood Duck    | 508.2     | PD006   | 623      | M   |
| Anatidae | Plant | Mallard      | 5747.2    | WN001   | 1058     | M   |
| Gaviidae | Fish  | Common Loon  | 277.4     | WN005   | 4400     | F   |

Figure 3 – Fully denormalized

| Family   | Diet  | Family   | SpeciesName  | Abundance | SpeciesName  | IndivID | Weight_g | Sex |
|----------|-------|----------|--------------|-----------|--------------|---------|----------|-----|
| Anatidae | Plant | Anatidae | Wood Duck    | 508.2     | Wood Duck    | LJ001   | 650      | M   |
| Gaviidae | Fish  | Anatidae | Mallard      | 5747.2    | Mallard      | LJ002   | 1050     | M   |
|          |       | Gaviidae | Pacific Loon | 10        | Pacific Loon | LJ003   | 4126     | M   |
|          |       | Gaviidae | Common Loon  | 277.4     | Common Loon  | LJ004   | 4002     | F   |
|          |       |          |              |           | Wood Duck    | PD001   | 605      | F   |
|          |       |          |              |           | Mallard      | PD004   | 1098     | M   |
|          |       |          |              |           | Wood Duck    | PD006   | 623      | M   |
|          |       |          |              |           | Mallard      | WN001   | 1058     | M   |
|          |       |          |              |           | Common Loon  | WN005   | 4400     | F   |

Figure 4 – Almost completely normalized

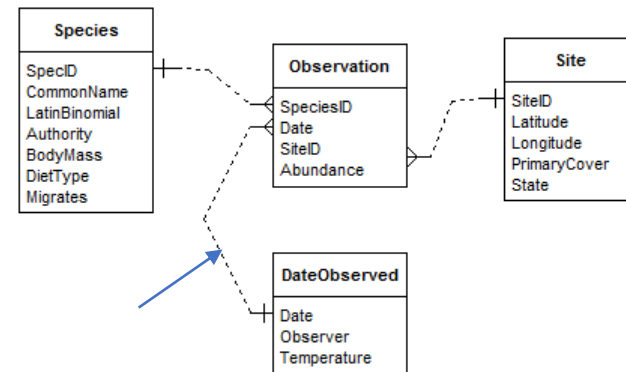
| Family   | Diet  | SpeciesName  | Abundance | SpeciesName  | IndivID | Weight_g | Sex |
|----------|-------|--------------|-----------|--------------|---------|----------|-----|
| Anatidae | Plant | Wood Duck    | 508.2     | Wood Duck    | LJ001   | 650      | M   |
| Anatidae | Plant | Mallard      | 5747.2    | Mallard      | LJ002   | 1050     | M   |
| Gaviidae | Fish  | Pacific Loon | 10        | Pacific Loon | LJ003   | 4126     | M   |
| Gaviidae | Fish  | Common Loon  | 277.4     | Common Loon  | LJ004   | 4002     | F   |
|          |       |              |           | Wood Duck    | PD001   | 605      | F   |
|          |       |              |           | Mallard      | PD004   | 1098     | M   |
|          |       |              |           | Wood Duck    | PD006   | 623      | M   |
|          |       |              |           | Mallard      | WN001   | 1058     | M   |
|          |       |              |           | Common Loon  | WN005   | 4400     | F   |

Figure 5 – partly normalized

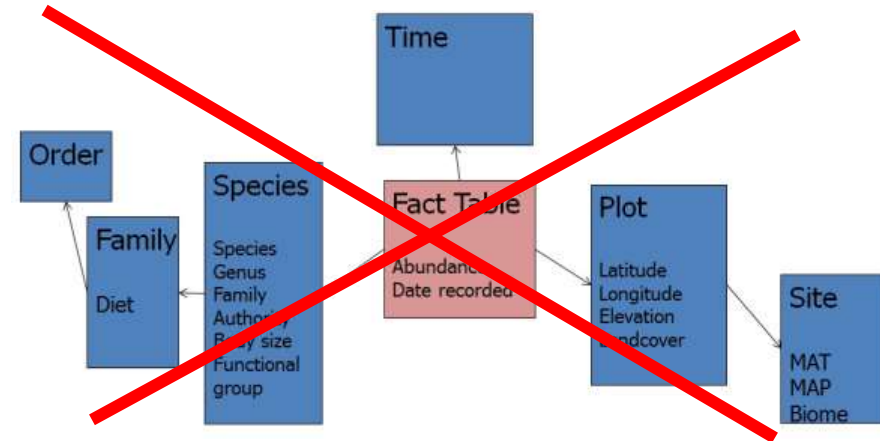


### 3 -Thou shalt partially (but no more and no less) normalize raw data into star schema(s)

- Star schema:

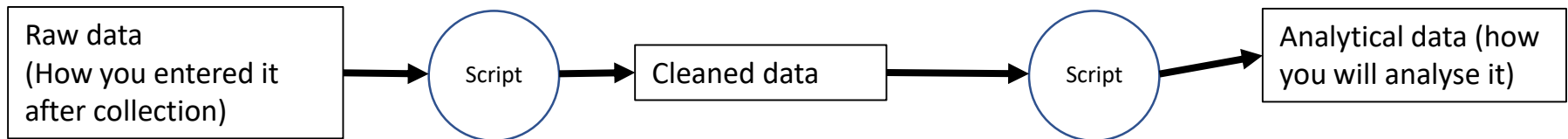


- Snowflake schema:  
(Fully normalized star  
schema, not  
recommended)





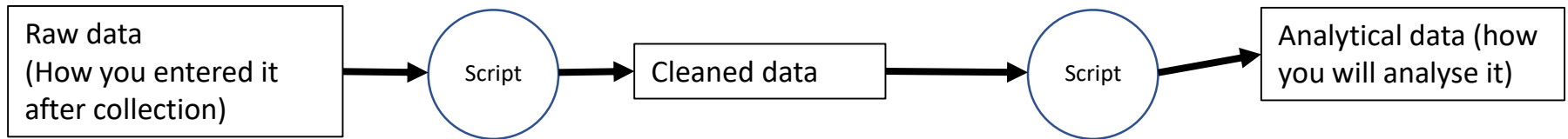
## 9 – Though shalt obsessively hand check the transformation from raw data to analytical data



- Do you have the correct number of rows and columns?
- Are there missing or double entries?
- Take a row in your analysis table and hand trace it back to the raw data. Can you reproduce that row from the raw data by hand calculations? If so take another entry and trace it back.
- Similarly scrutinize your analysis table. Do you have a species called “NA” (or a time period or site)? How did that sneak in there? Is an endangered species showing up as your most common species?
- Did your joins work so that a species is lined up with the right family? Really poke and prod your data.



## 9 – Though shalt obsessively hand check the transformation from raw data to analytical data



- Unit tests

- `if(any(is.na(data$species_name))) warning(«missing species name»)`
  - `if(any(!is.integer(data$abundance))) warning(«wrong abundance value»)`

- R packages Checkmate and Testthat



# Some Simple Guidelines for Effective Data Management (Borer 2009)

*Van Klink*

- **3. Store data in nonproprietary hardware formats**
- They mean: don't use a format that you need a special machine for to read it
- I would add: use formats that do not require paywalled software (such as Microsoft products), but a format that is most likely to be readable 100 years into the future: tab or comma delimited text files (.txt or .csv, avoiding special characters [ä,é,€ etc])
- "As hard as it is to believe today, we can foresee the day when CD-ROMs might be difficult to read. [...] At various times, it is also advisable to create additional copies of your data that are off-line (not on the Internet), using the most popular medium of the day. As of 2008, this is probably the DVD..."



# Some Simple Guidelines for Effective Data Management (Borer 2009)

*Van Klink*

- 5. Use descriptive names for your data files.
- 7. Use plain ASCII text for your file names, variable names, and data values.
  - I add: use identical column names for the columns that will link your different tables by this column name (e.g. “Plot\_ID” in both)
- 11. Record full information about taxonomic names.
- 12. Record full dates, using standardized formats.
- 13. Always maintain effective metadata.





FILE: 'FRESHWATER  
EXPERIMENT'



OPEN THE FILE (IN EXCEL  
OR R), BROWSE  
THROUGH IT,



COMPARE TO THE  
COMMANDMENTS OF  
MCGILL AND ME: WHAT  
ARE THEY DOING  
WRONG?

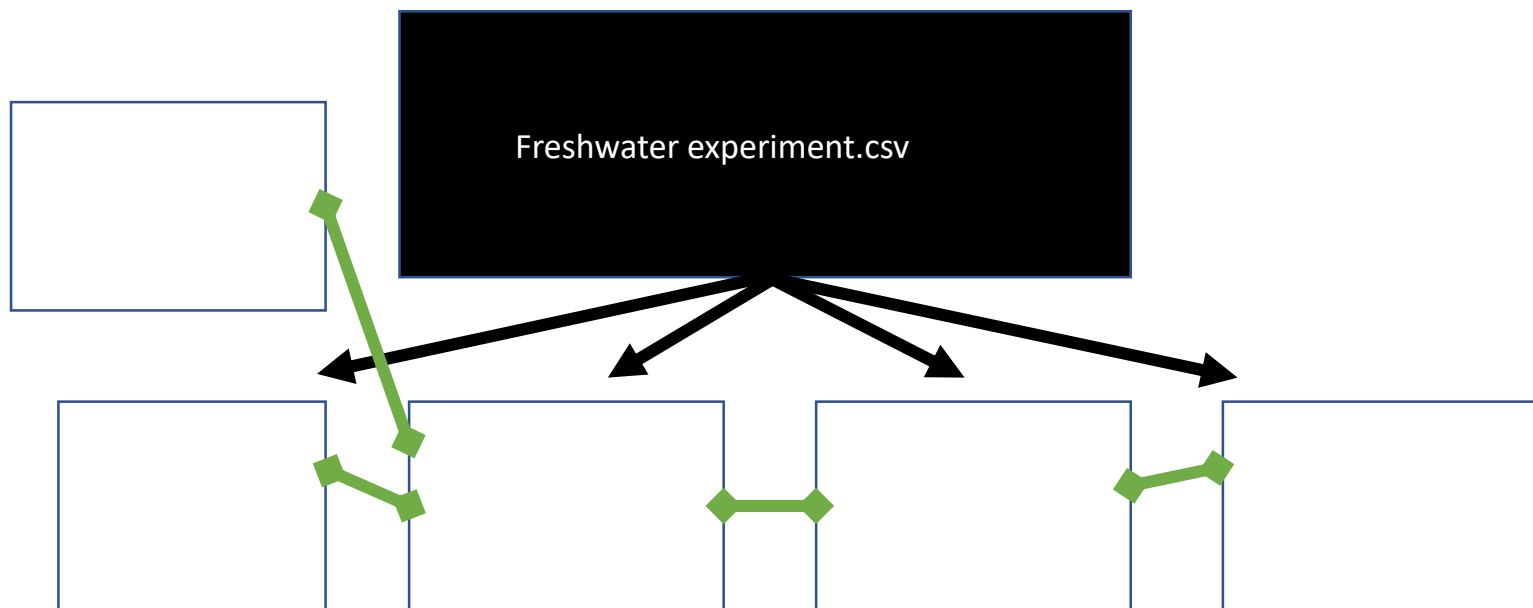


HOW COULD IT BE DONE  
BETTER?





- Study the columns. Which columns contain redundant information that is better stored in separate tables? (Hint: typically, data that remain true (for the duration of the study), go together into the same table)
- Which information goes into which table? How would you call the table?
- Draw your star-schema (which doesn't need to be star shaped! It can be linear) for storing this dataset on the white board. If you spot mistakes or missing information, take notes!
- Bonus: can you think of information that is currently missing from the dataset, but would be useful? In which table would you store this? Hint: traits
- Bonus 2: did you spot any mistakes / bad practice?



# Resources and tips:

*Van Klink*

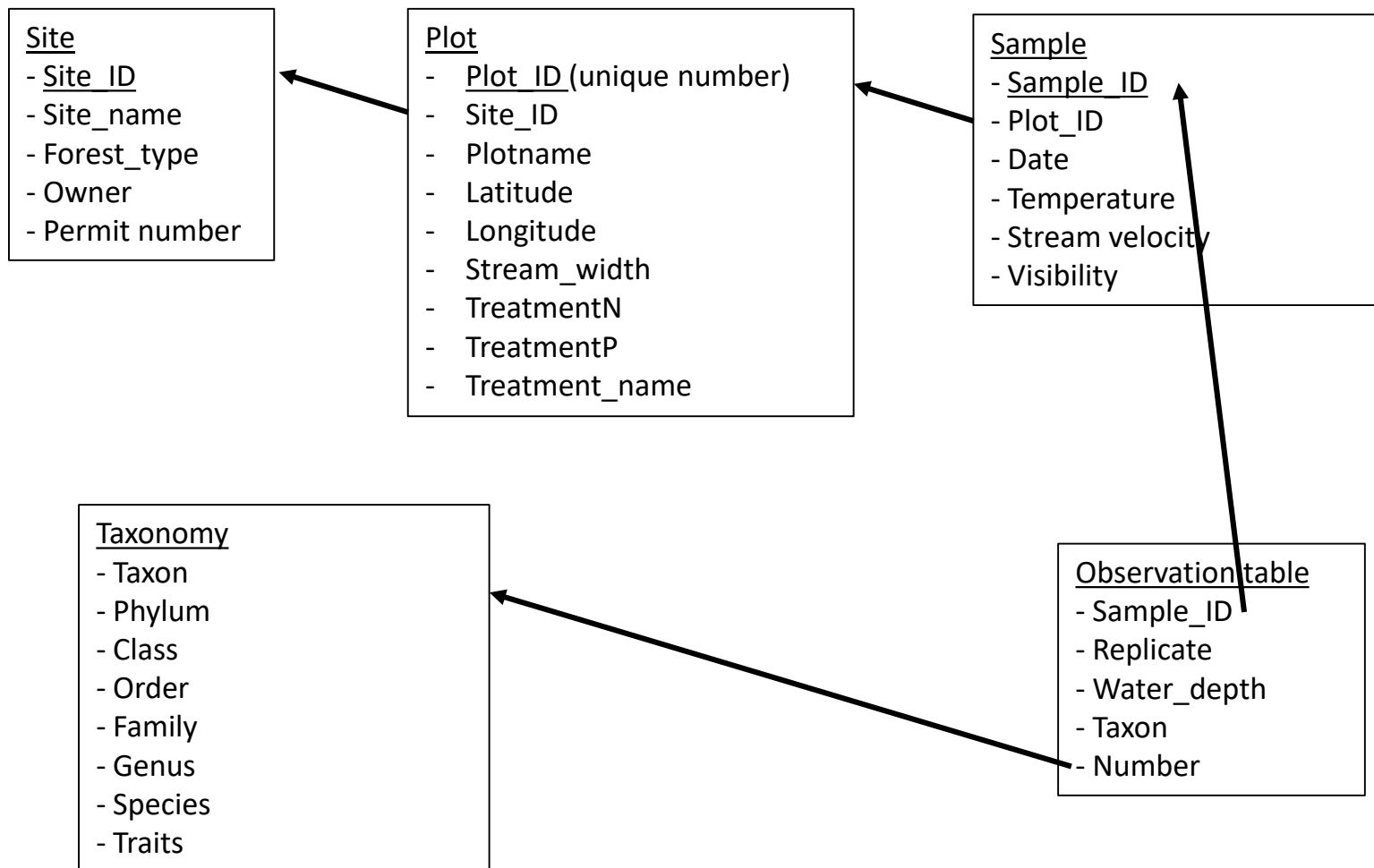
- Brian McGill blogpost + script
- Publications by Borer (2009) and Costello (2014)
- Tidyverse: <https://subhayo.wordpress.com/2017/12/16/data-manipulation-of-star-wars-characters-using-dplyr-and-tidyr/>
- Pivot tables:

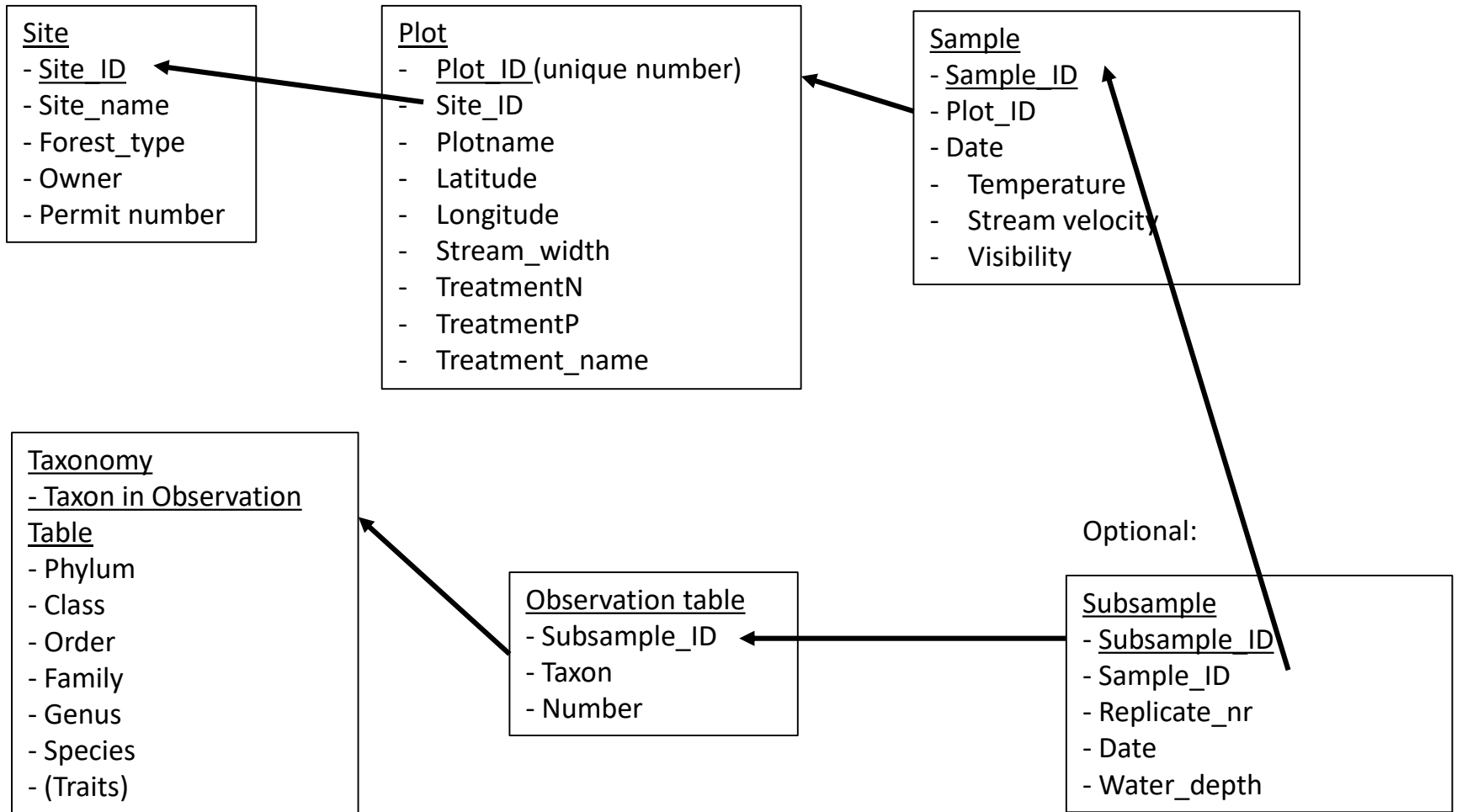
Reshape2 library: dcast – has summary functions like sum, mean, sd,  
Tidyverse::pivot\_wider. Excel can also do this: insert: pivot table.

Data extraction examples: [www.github.com/chase-lab/biotimex](http://www.github.com/chase-lab/biotimex)

Data extraction tutorial: <https://psyteachr.github.io/reprores-v2/>







# Not recommended:

*Van Klink*

## Classes

- Phylum
- Class

## Orders

- Class
- Order

## Families

- Order
- Family

## Genera

- Family
- Genus

## Species

- Genus
- Species

## Trait

- Taxon
- Trait\_X
- Trait\_value



# Open science



# What is the definition of Open?

Open means anyone can **freely access, use, modify, and share** for any purpose  
(subject, at most, to requirements that preserve provenance and openness)

<https://opendefinition.org>

An Open work must:

1. Possess an Open licence or be in the public domain
2. Be accessible at reasonable cost (but doesn't have to be online)
3. Be machine readable
4. Open format – can be processed by  $\geq 1$  open source software



# The FAIR data principles

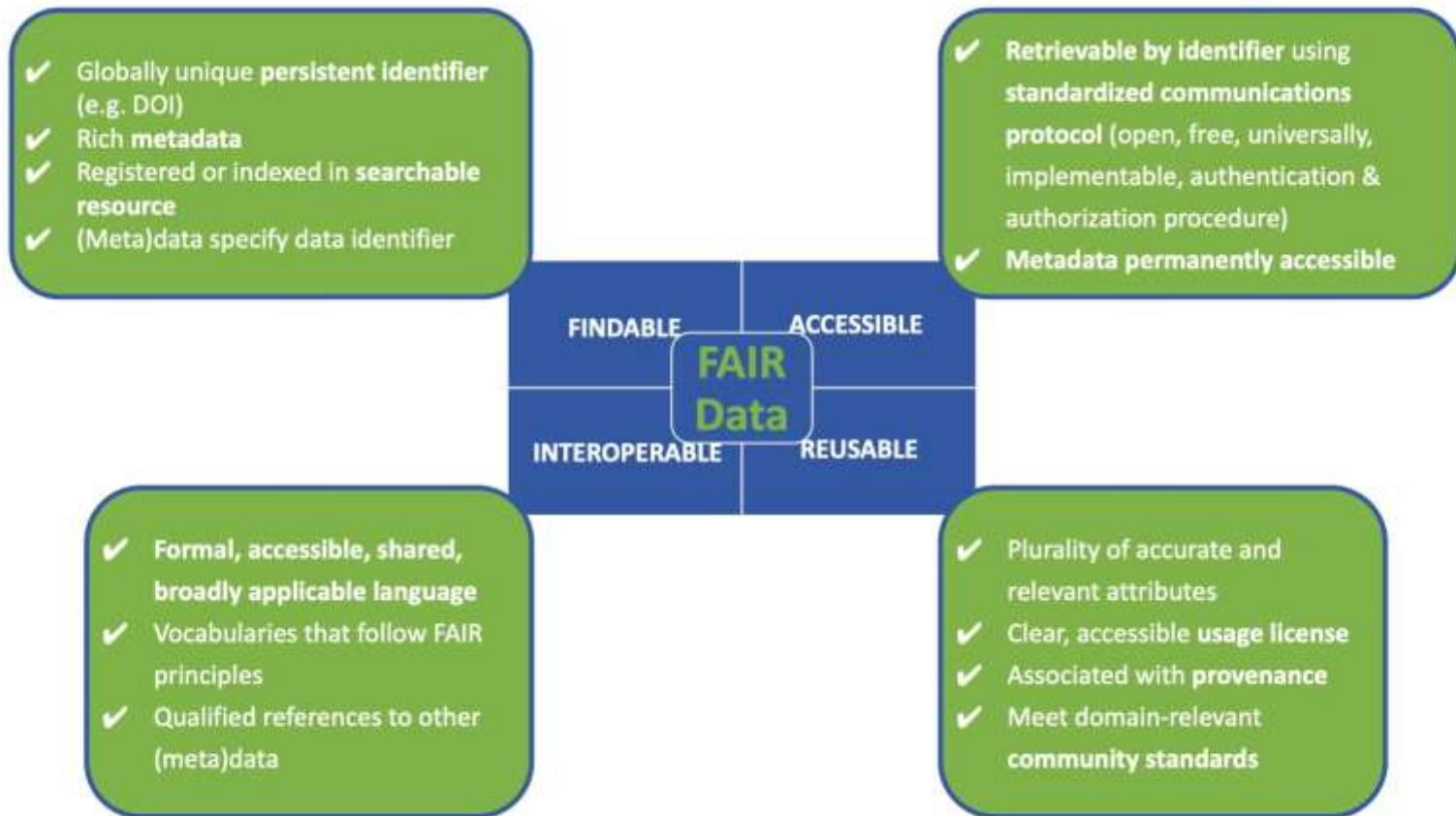


- Published in 2016
- Basic requirements for reusable data
- Increasingly important in science
- Part of the DFG 'Guidelines for Safeguarding Good Scientific Practice'
- Substantial part of DMPs in H2020 projects

Note that FAIR  $\neq$  Open - we should strive for FAIR/O



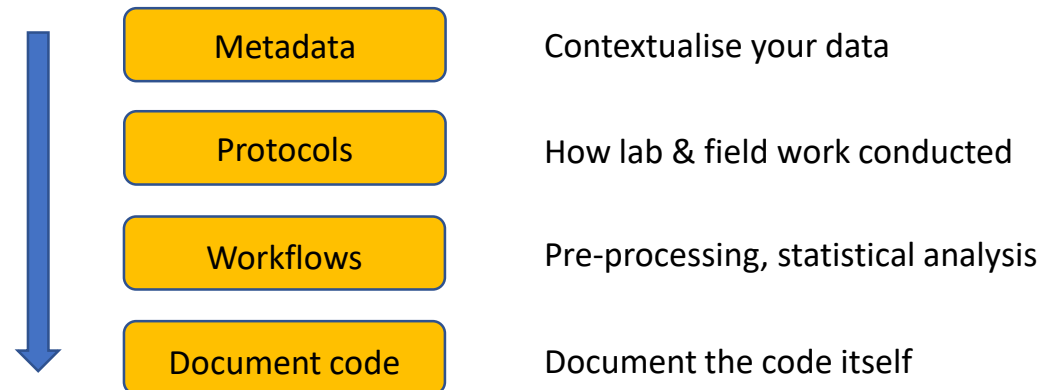




Source: Wilkinson MD, Dumontier M, Aalbersberg JJ, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, Issue 3, 10.1038/sdata.2016.18



# Documenting your work



Do this from the start, and continue doing it throughout



# What counts as (research) data?



- Measurement data (observational or lab values)
- Audiovisual information
- Remote sensing
- Observational data or surveys
- Interviews or questionnaires
- Written texts
- Physical objects (archaeological, tissue samples)
- Software and simulations



# What is metadata?

'Data about data'

Basic structured information about your data so that others (or your Future Self) can understand & use the data *without needing additional information*

It should answer the 6 questions:



Metadata is an important part of making your data **F-A-I-R**

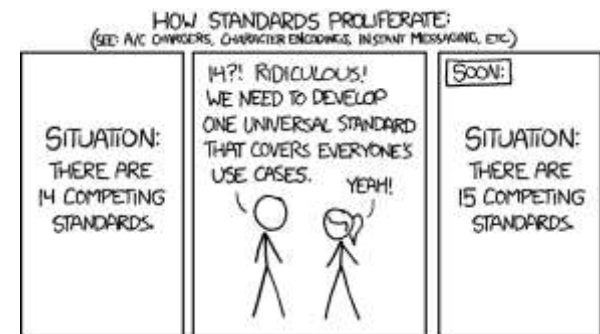


Image: XKCD, CC BY-NC



# Where to publish your code?

## Just as with datasets

- Don't leave in Supplementary Info of paper
- Get a DOI & make your code citable
- Repository should guarantee for  $\geq 10$  years
- Choose a license



GitHub is not an 'archive'

Table 1. Comparison of Common Resources (Zenodo, Figshare, Dryad Digital Repository, PANGAEA Data Publisher, GitHub, and Bitbucket) Used for Archiving Code and Data<sup>a</sup>

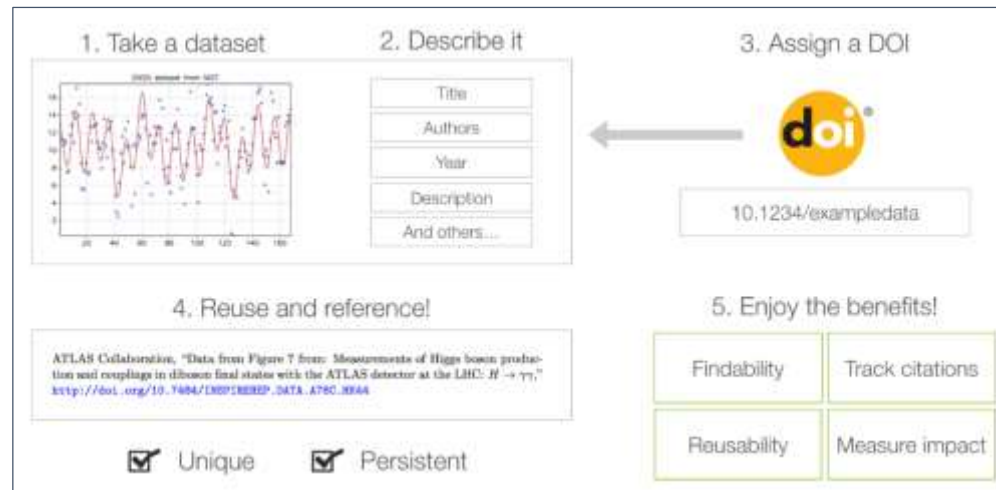
|                    | Zenodo           | Figshare         | Dryad            | PANGAEA          | GitHub and Bitbucket | Supplementary Material |
|--------------------|------------------|------------------|------------------|------------------|----------------------|------------------------|
| Default License    | Flexible         | MIT              | CC0              | CC-BY            | Flexible             | None                   |
| Long-term          | Yes <sup>b</sup> | Yes <sup>b</sup> | Yes <sup>b</sup> | Yes <sup>b</sup> | No                   | Yes <sup>b</sup>       |
| Assigns DOI        | Yes              | Yes              | Yes              | Yes              | No                   | No                     |
| Code Search Option | Yes              | Yes              | No               | No               | Yes                  | No                     |
| Upload from GitHub | Yes              | No               | No               | No               | –                    | No                     |
| Cost to Author     | None             | None             | Possible         | None             | None                 | None                   |

Mislan et al. 2016 *Trends Ecol Evol* 31: 4-7



## Assign a DOI (Digital Object Identifier)

- Globally unique, alphanumeric string assigned by a registration agency
- When citing, write full URL: <https://doi.org/10.1234/EXAMPLEDATA>
- Must resolve to a landing page, containing metadata about the resource



[doi.datacite.org](https://doi.org/10.1234/EXAMPLEDATA)

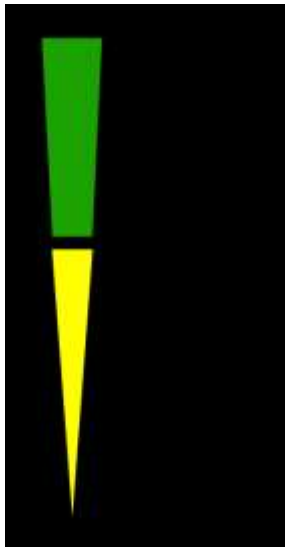


If a dataset/code does not specify a license, then noone can reuse it!  
... even if you made the resource easily accessible online



# Specify a licence

Creative Commons licenses - for **datasets**, images, etc



[www.creativecommons.org](http://www.creativecommons.org)

## Advantages of CC licenses

- Standardized license text → reduces effort & creates legal certainty
- No transfer of exclusive exploitation rights
- Human & machine readable
- Internationally accepted
- Can be combined



CC0: public domain, no restrictions  
CC-BY: credit must be given to the creator  
CC-SA: adaptations must be shared under the same terms  
CC-ND: no derivatives or adaptations of the work permitted  
CC-NC: only noncommercial uses of the work permitted

Some repositories use the same license for all datasets – Dryad uses CC0.  
Just because you didn't actively select a license doesn't mean that there is no license.

DFG & the EU recommend publishing research data under the **CC-BY 4.0** license





## Choose open file formats

- Machine readable - plain text rather than binary file format
- Non-proprietary - readable by at least 1 free software package

| File Type       | Recommendation                                             | Do not use       |
|-----------------|------------------------------------------------------------|------------------|
| Tables          | CSV, TSV, SPSS portable                                    | Excel            |
| Text            | TXT, MD, ODT, HTML, RTF, PDF/A only if layout is important | Word, Powerpoint |
| Multimedia      | Container: MP4, Ogg<br>Codec: Theora, Dirac, FLAC          | QuickTime, H264  |
| Pictures        | TIFF, JPEG2000, PNG, SVG                                   | GIF, JPG         |
| Structured Data | XML, RDF, JSON                                             | RDBMS            |

Helbig 2017

See UK Data Service guidance on recommended formats  
DataONE Best Practices for file formats



# Specify the character encoding

... first when **saving** your files  
... and **state** it when publishing

## I (guess and) specify 'Windows Latin 1' encoding

The experiment was carried out in Research Arboretum Großpöсна (51°15'41"N, 12°29'55"E), with the following number of leaves per tree species: C. betulus – 45 leaves, Q. robur – 25 leaves, T. cordata 30 leaves.

## Opened with default encoding used by TextEdit on Mac OS

The experiment was carried out in Research Arboretum Großpöсна (51°15'41"N, 12°29'55"E), with the following number of leaves per tree species: C. betulus ħ 45 leaves, Q. robur ħ 25 leaves, T. cordata 30 leaves.

## Opened with the default encoding used by MS Word 2011 on Mac OS

The experiment was carried out in Research Arboretum Großpöсна (51°15'41"N, 12°29'55"E), with the following number of leaves per tree species: C. betulus ů 45 leaves, Q. robur ů 25 leaves, T. cordata 30 leaves.

Recommendation: use **UTF-8** (ASCII characters are a subset)  
Only use UTF-16 if you need to (takes more space)



# Break



# **Matching Species Names Across Biodiversity Databases:**

*Sources, tools, pitfalls and best practices for taxonomic harmonization*

Matthias Grenié, **Emilio Berti**, Juan Carvajal-Quintero, **Alban Sagouis**, Marten Winter







## Dataset A

Sp1  
Sp2  
Sp3  
Sp4  
—  
—

## Dataset B

—  
Sp2  
Sp3  
Sp4  
—  
Sp6



## Dataset A



Sp1

Sp2

Sp3

Sp4

—

—

## Dataset B

—

Sp2

Sp3

Sp4

—

Sp6

*Acer monspessolanum*





## Dataset A



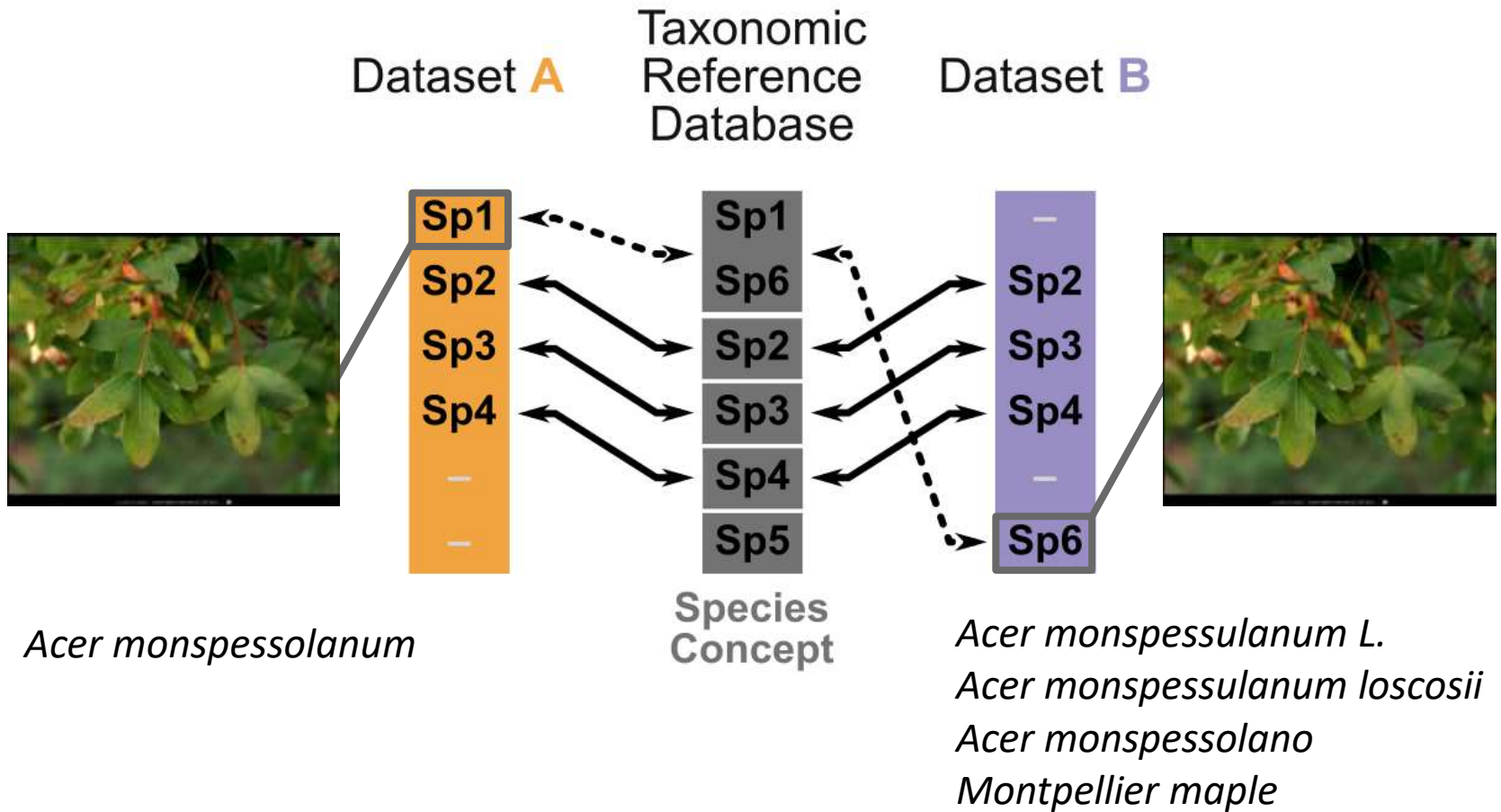
*Acer monspessulanum*

## Dataset B



*Acer monspessulanum* L.  
*Acer monspessulanum loscosii*  
*Acer monspessolano*  
Montpellier maple







**Is there one, best way to harmonize taxonomy?**





## Is there one, best way to harmonize taxonomy?

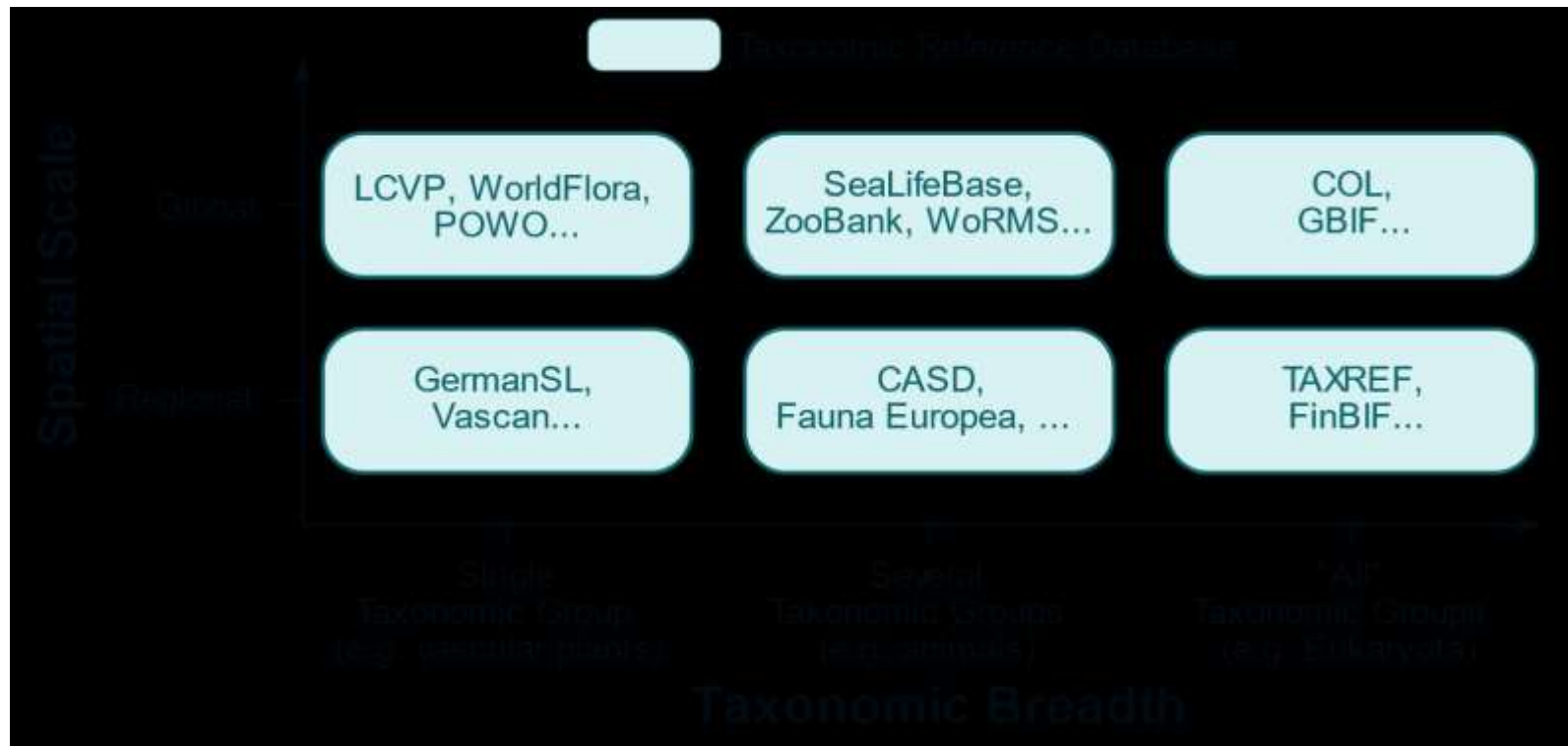
What are the available resources?

- Databases
- Tools (R packages)



45 databases

## A typology of taxonomic databases



61 packages

## Available R packages

| Category       | Packages               |
|----------------|------------------------|
| Infrastructure | taxa, taxlist, taxview |
|                |                        |
|                |                        |
|                |                        |
|                |                        |

3



61 packages

## Available R packages

| Category                 | Packages                                                                                                                                                                                                                                                                                                                                      |    |
|--------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Infrastructure           | taxa, taxlist, taxview                                                                                                                                                                                                                                                                                                                        | 3  |
| Database access (online) | algaeClassify, AmphiNom, dyntaxa, finbif, flora, mammals, natserv, neotoma2, paleobioDB, plantlist, rcol, rebird, rentrez, <b>rfishbase</b> , <b>rgbif</b> , ritis, Rocc, rotl, rredlist, rtaxref, SP2000, <b>taxize</b> , taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twm, wikipata, worms, worrms, zbank, kewr | 35 |
|                          |                                                                                                                                                                                                                                                                                                                                               |    |
|                          |                                                                                                                                                                                                                                                                                                                                               |    |
|                          |                                                                                                                                                                                                                                                                                                                                               |    |



61 packages

## Available R packages

| Category                  | Packages                                                                                                                                                                                                                                                                                                                                      |    |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Infrastructure            | taxa, taxlist, taxview                                                                                                                                                                                                                                                                                                                        | 3  |
| Database access (online)  | algaeClassify, AmphiNom, dyntaxa, finbif, flora, mammals, natserv, neotoma2, paleobioDB, plantlist, rcol, rebird, rentrez, <b>rfishbase</b> , <b>rgbif</b> , ritis, Rocc, rotl, rredlist, rtaxref, SP2000, <b>taxize</b> , taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twm, wikipata, worms, worrms, zbank, kewr | 35 |
| Database access (offline) | <b>lcvplants</b> , ncbi, splister, taxadb, taxalight, taxastand, taxizedb, taxonlookup, <b>vegdata</b> , WorldFlora                                                                                                                                                                                                                           | 10 |
|                           |                                                                                                                                                                                                                                                                                                                                               |    |
|                           |                                                                                                                                                                                                                                                                                                                                               |    |





61 packages

## Available R packages

| Category                  | Packages                                                                                                                                                                                                                                                                                                                                      |    |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Infrastructure            | taxa, taxlist, taxview                                                                                                                                                                                                                                                                                                                        | 3  |
| Database access (online)  | algaeClassify, AmphiNom, dyntaxa, finbif, flora, mammals, natserv, neotoma2, paleobioDB, plantlist, rcol, rebird, rentrez, <b>rfishbase</b> , <b>rgbif</b> , ritis, Rocc, rotl, rredlist, rtaxref, SP2000, <b>taxize</b> , taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twn, wikipata, worms, worrms, zbank, kewr | 35 |
| Database access (offline) | <b>lcvplants</b> , ncbi, splister, taxadb, taxalight, taxastand, taxizedb, taxonlookup, <b>vegdata</b> , WorldFlora                                                                                                                                                                                                                           | 10 |
| Data wrangling            | metacoder, monographR, <b>rgnparser</b> , splister, taxastand, taxreturn, taxspell, traitdatafrom, vegdata, vegtable, yatah                                                                                                                                                                                                                   | 11 |
|                           |                                                                                                                                                                                                                                                                                                                                               |    |



61 packages

## Available R packages

| Category                  | Packages                                                                                                                                                                                                                                                                                                                                      |    |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Infrastructure            | taxa, taxlist, taxview                                                                                                                                                                                                                                                                                                                        | 3  |
| Database access (online)  | algaeClassify, AmphiNom, dyntaxa, finbif, flora, mammals, natserv, neotoma2, paleobioDB, plantlist, rcol, rebird, rentrez, <b>rfishbase</b> , <b>rgbif</b> , ritis, Rocc, rotl, rredlist, rtaxref, SP2000, <b>taxize</b> , taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twn, wikipata, worms, worrms, zbank, kewr | 35 |
| Database access (offline) | <b>lcvplants</b> , ncbi, splister, taxadb, taxalight, taxastand, taxizedb, taxonlookup, <b>vegdata</b> , WorldFlora                                                                                                                                                                                                                           | 10 |
| Data wrangling            | metacoder, monographR, <b>rgnparser</b> , splister, taxastand, taxreturn, taxspell, traitdatafrom, vegdata, vegtable, yatah                                                                                                                                                                                                                   | 11 |
| Data visualization        | metacoder, taxview                                                                                                                                                                                                                                                                                                                            | 2  |



45 Databases and relationships  
among them

35 R packages accessing  
databases

61 R packages and relationships  
between them





taxharmonizeexplorer [Description](#) [Network](#) [Help](#)

**Selected Node Information**

**Name:** Tropicos  
**Full Name:** Tropicos  
**Type:** database  
**Taxonomic Group:** Vascular plants  
**Spatial Scale:** Global  
**Taxonomic Breadth:** Small  
**URL:** <https://tropicos.org/name/Search>

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

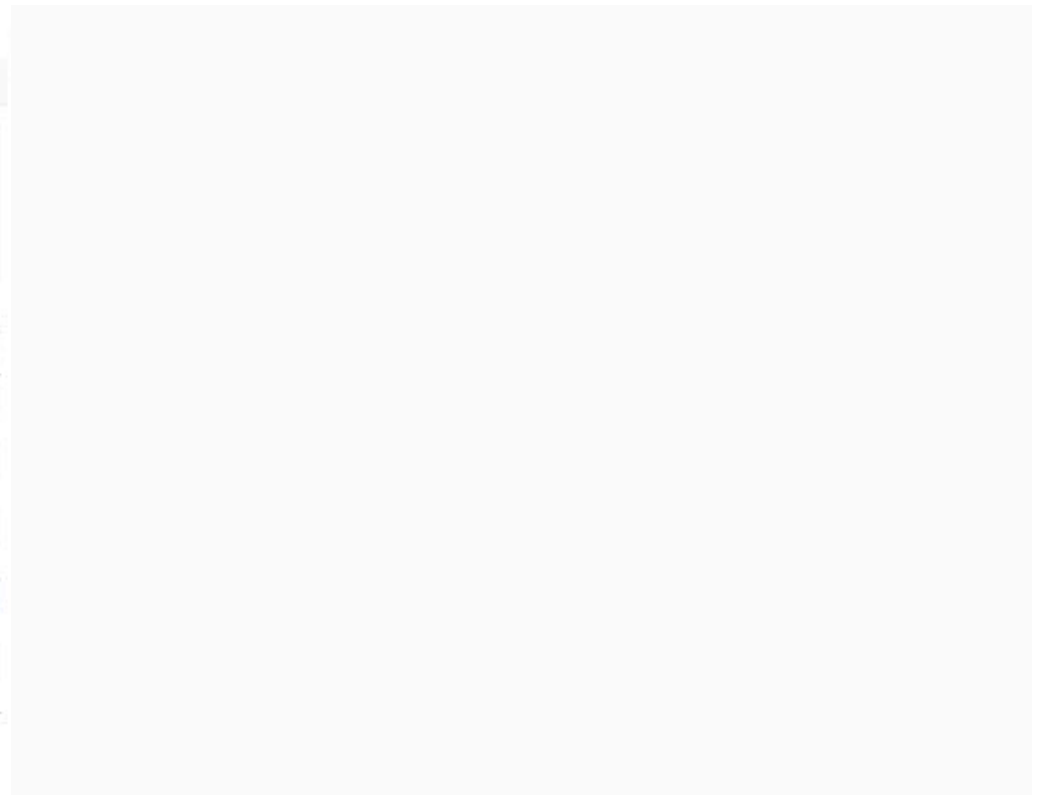
Show  entries

Search:

|    | Name        | Type     | Tax. Group      |
|----|-------------|----------|-----------------|
| 33 | taxlist     | package  | land plants     |
| 34 | taxonlookup | package  | land plants     |
| 36 | Taxonstand  | package  | land plants     |
| 88 | TNRS        | database | Plants          |
| 40 | tpi         | package  | plants          |
| 62 | TPL         | database | Vascular plants |
| 74 | Tropicos    | database | Vascular plants |
| 89 | USDA        | database | Vascular plants |
| 80 | Vascan      | database | Vascular plants |
| 43 | vegdata     | package  | land plants     |

Showing 11 to 20 of 24 entries (filtered from 98 total entries)

Previous    Next





taxharmonizeexplorer

Description

Network

Help

## Selected Node Information

Name: Tropicos

Full Name: Tropicos

Type: database

Taxonomic Group: Vascular plants

Spatial Scale: Global

Taxonomic Breadth: Small

URL: <https://tropicos.org/name/Search>

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

Show 10 entries

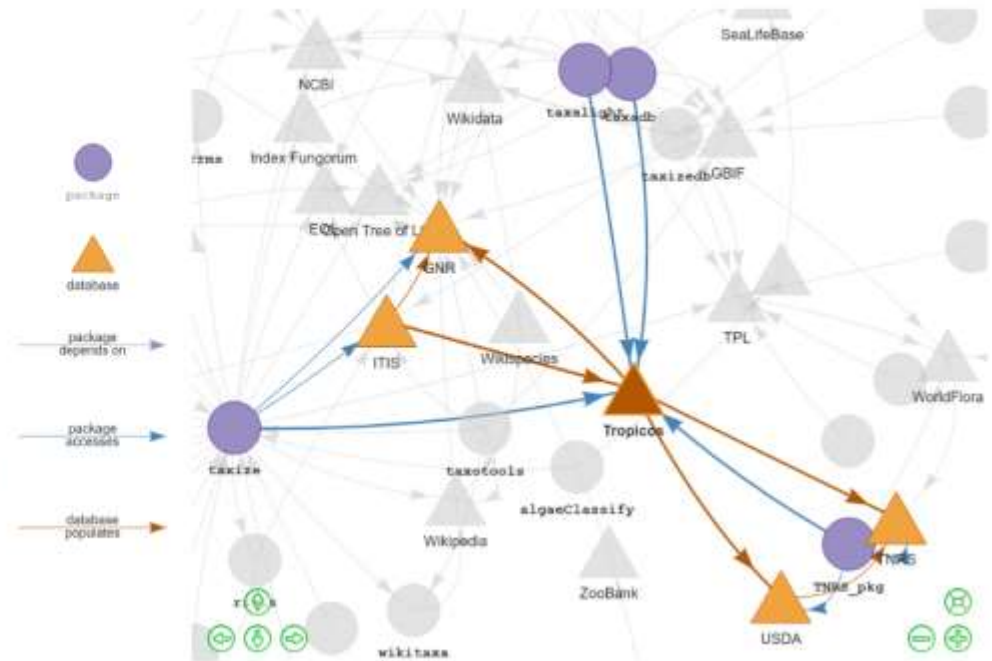
Search: 

|    | Name        | Type     | Tax. Group      |
|----|-------------|----------|-----------------|
| 33 | taxit       | package  | land plants     |
| 34 | taxonlookup | package  | land plants     |
| 36 | Taxonstand  | package  | land plants     |
| 88 | TNRS        | database | plants          |
| 40 | tpl         | package  | plants          |
| 62 | TPL         | database | Vascular plants |
| 74 | Tropicos    | database | Vascular plants |
| 89 | USDA        | database | Vascular plants |
| 80 | Vascan      | database | Vascular plants |
| 43 | vegdata     | package  | land plants     |

Showing 11 to 20 of 24 entries (filtered from 98 total entries)

Previous 1 2 3 Next

## Relationships between taxonomic R packages and databases





## Is there one, best way to harmonize taxonomy?

What are the available resources?

- Databases
- Tools (R packages)

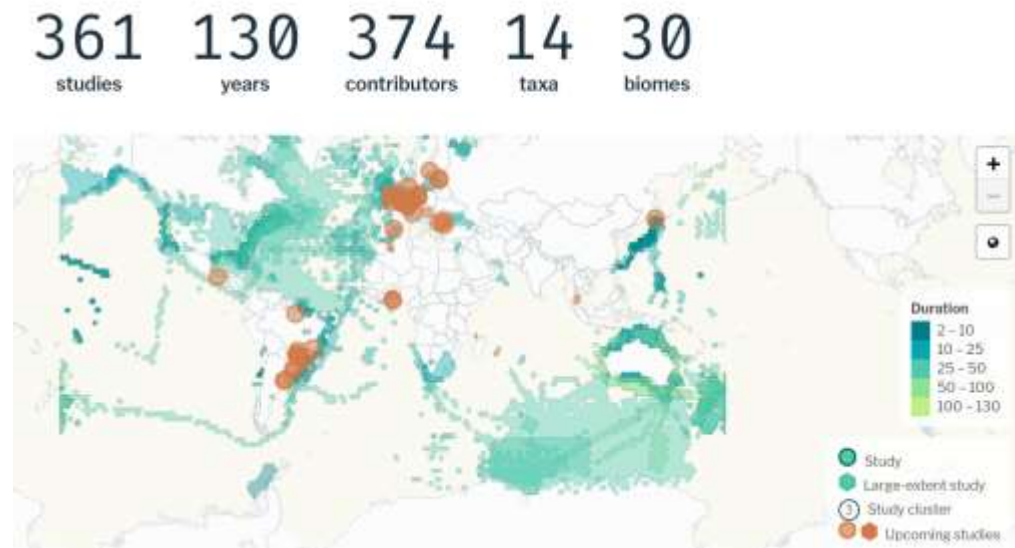
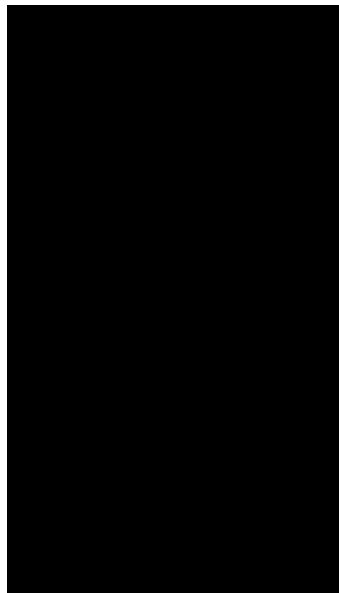
Which workflows are most suitable?

- Accurate
- Easy to implement
- Matching most names



## Harmonize BioTIME database

Global database of assemblage time series



Only birds, fishes, and vascular plants



Raw list

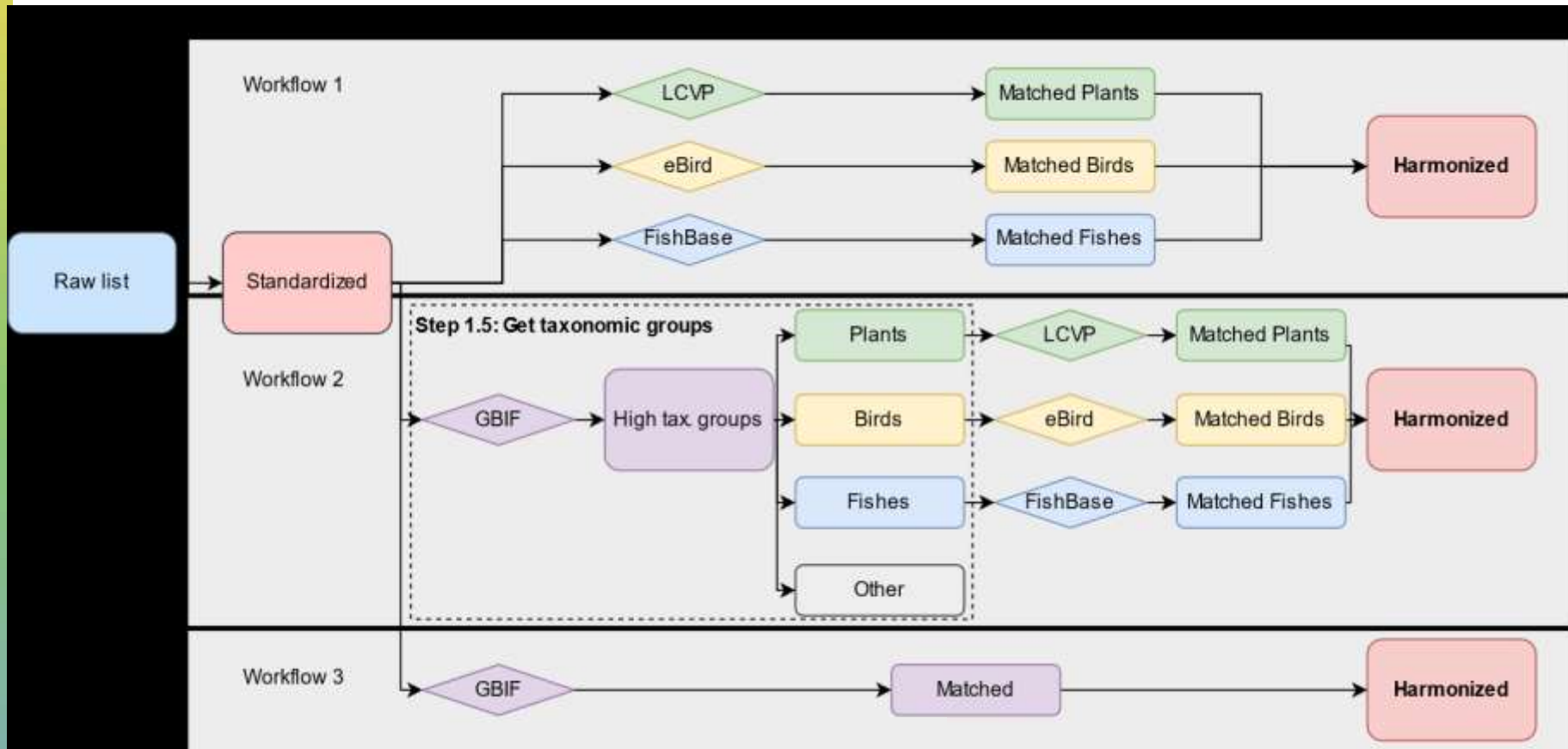




Raw list

gn\_parse\_tidy() from package rgnparser v.0.2.0





More in the paper:

- Recommendations to users, developers and database managers
- Warning on Outdated online resources
  - *The Plant List* (+2013)
  - *Global names Index* (2018) / *Resolver* (2021) / *Verifier*
- The double-edged sword of “fuzzy matching”

Shiny app: <https://mgrenie.shinyapps.io/taxtool-selector/>

Pre-print: <https://doi.org/10.32942/osf.io/e3qnz>



# Thank you

*Berti, Sagouis*

