Data management - Data structuring

Tips for data structuring

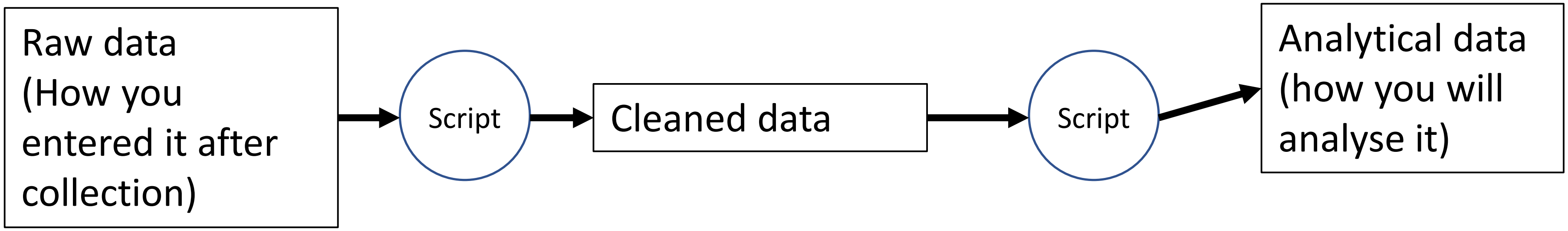
McGill‘s 10 commandments

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



Raw data (How you entered it after collection)

Script

Cleaned data

Script

Analytical data (how you will analyse it)

-Each observation is entered
-Minimize chance for errors
-Clean
-Well-structured
-Flexible format

- Errors fixed
- Some data excluded

- Only certain data selected?
- Some observations aggregated?
- Rows and columns ordered in the structure needed

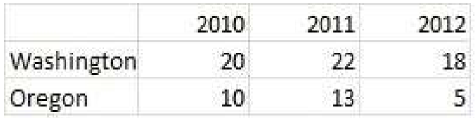
2 – Thou shalt create raw data in instance-row, variable-column format

Van Klink

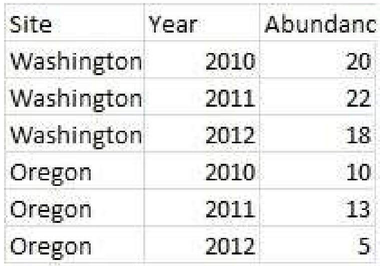
2 – Thou shalt create raw data in instance-row, variable-column format

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

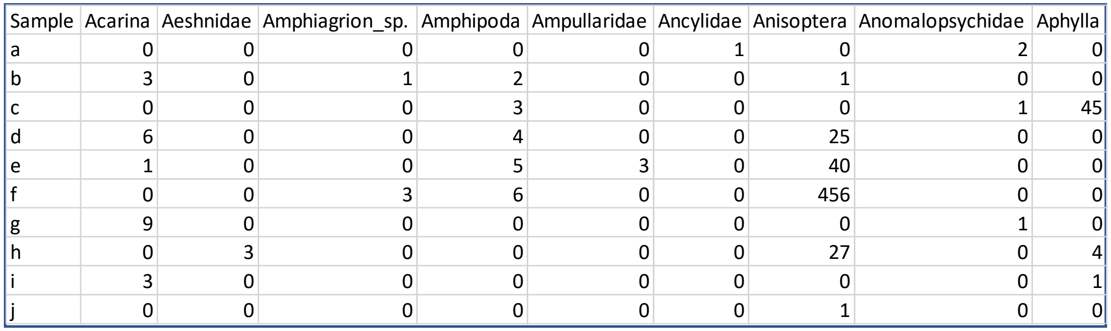


Site	Year	Abundanc
Washington	2010	20
Washington	2011	22
Washington	2012	18
Oregon	2010	10
Oregon	2011	13
Oregon	2012	5

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

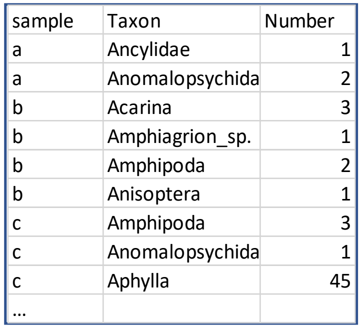
Van Klink

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:

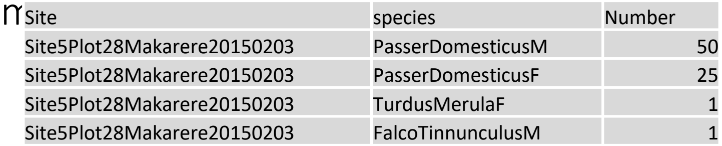


sample	Taxon	Number
a	Ancylidae	1
a	Anomalopsychida	2
b	Acarina	3
b	Amphiagrion_sp.	1
b	Amphipoda	2
b	Anisoptera	1
c	Amphipoda	3
c	Anomalopsychida	1
c	Aphylla	45
…		

More commandments from Roel

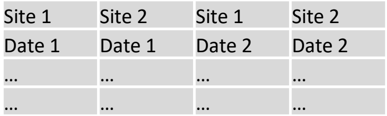
Van Klink

- 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 Exam



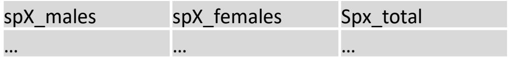
Site	species	Number
Site5Plot28Makarere20150203	PasserDomesticusM	50
Site5Plot28Makarere20150203	PasserDomesticusF	25
Site5Plot28Makarere20150203	TurdusMerulaF	1
Site5Plot28Makarere20150203	FalcoTinnunculusM	1

- 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

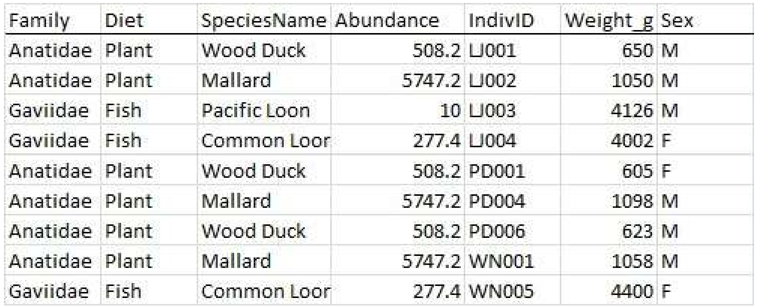


spX_males	spX_females	Spx_total
…	…	…

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

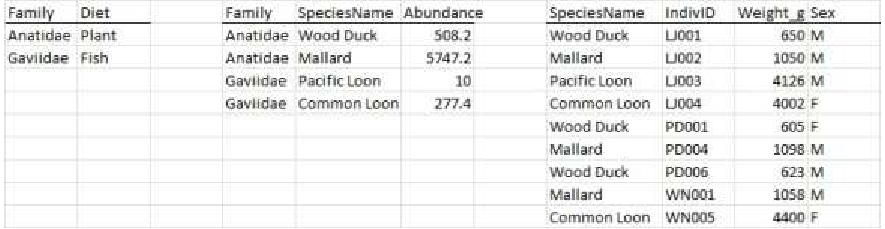
Van Klink

- 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 Loor	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 Loor	277.4	WN005	4400	F

Figure 3 – Fully denormalized

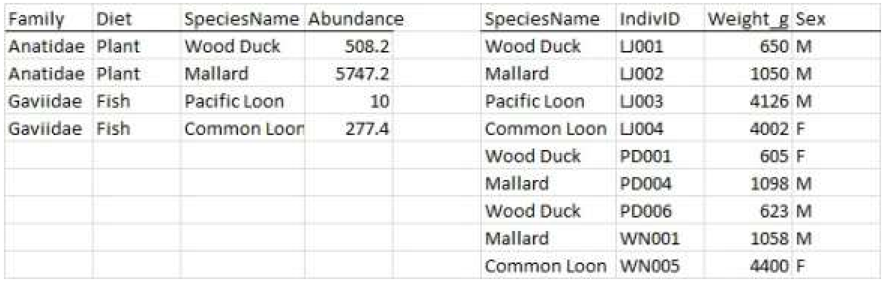


Family	Diet
Anatidae	Plant
Gaviidae	Fish

Family	SpeciesName	Abundance
Anatidae	Wood Duck	508.2
Anatidae	Mallard	5747.2
Gaviidae	Pacific Loon	10
Gaviidae	Common Loon	277.4

SpeciesName	IndivID	Weight_g	Sex
Wood Duck	LJ001	650	M
Mallard	LJ002	1050	M
Pacific Loon	LJ003	4126	M
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
Anatidae	Plant	Wood Duck	508.2
Anatidae	Plant	Mallard	5747.2
Gaviidae	Fish	Pacific Loon	10
Gaviidae	Fish	Common Loor	277.4

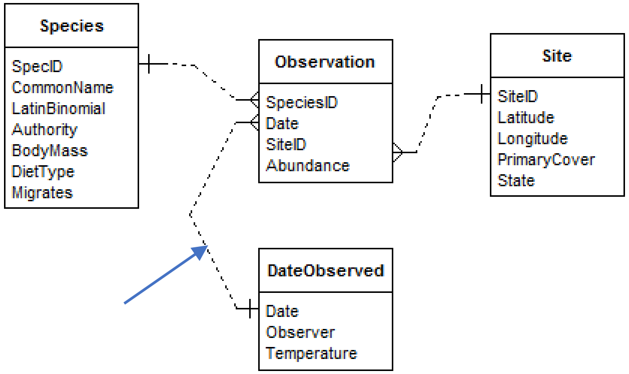
SpeciesName	IndivID	Weight_g	Sex
Wood Duck	LJ001	650	M
Mallard	LJ002	1050	M
Pacific Loon	LJ003	4126	M
Common Loon	LJ004	4002	F
Wood Duck	PD001	605	F
Mallard	PD004	1098	M
Wood Duck	PD006	623	M
Mallard	WN001	1058	M
Common Loon	WN005	4400	F

Figure 5 – partly normalized

Van Klink

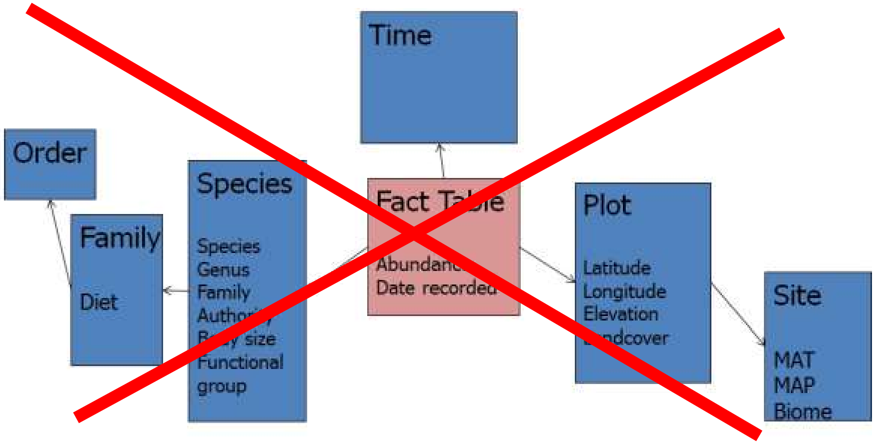
3 -Thou shalt partially (but no more and no less) normalize raw data into star schema(s)

- Star schema:

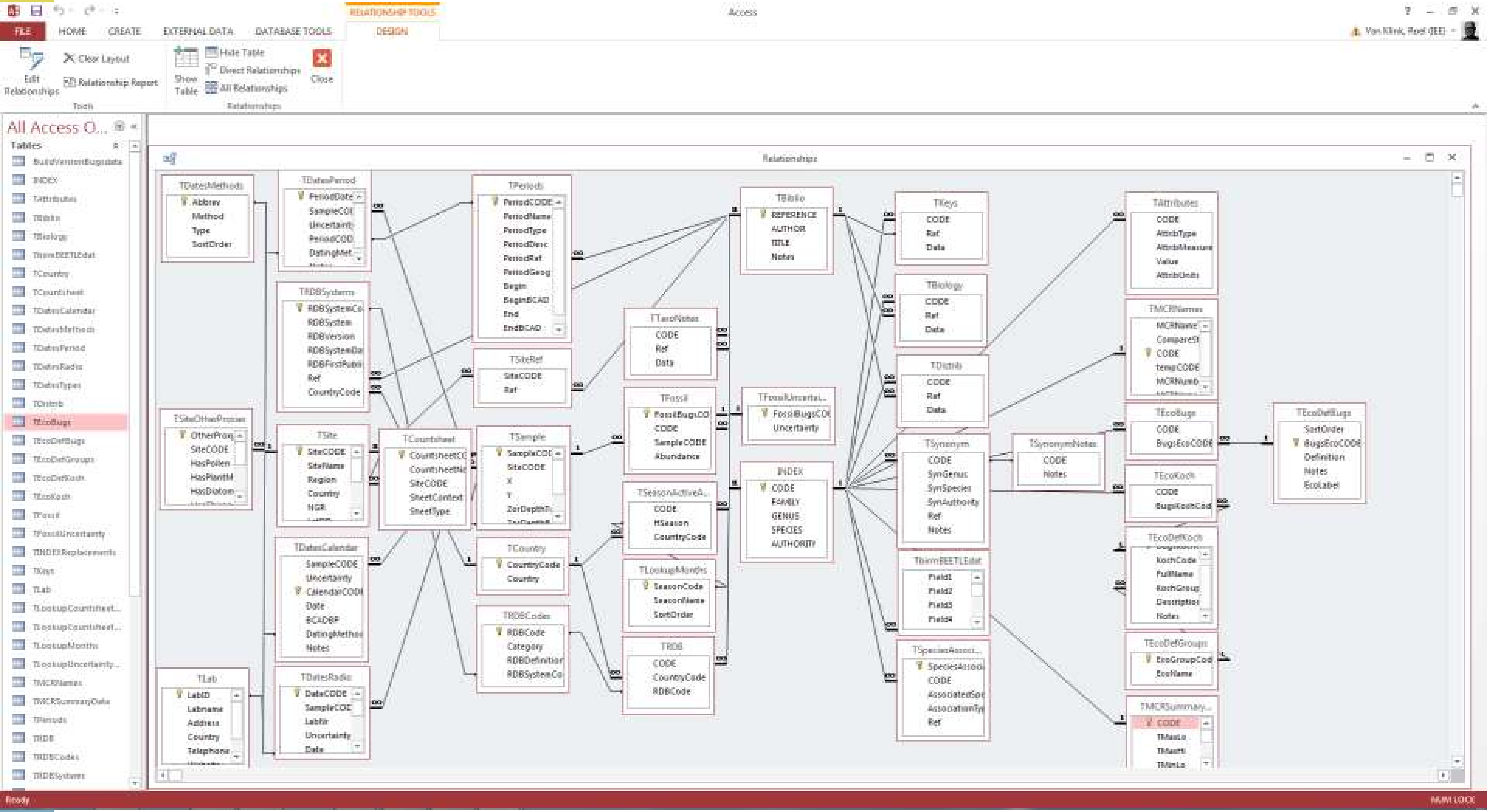


- Snowflake schema:

  (Fully normalized star schema, not recommended)

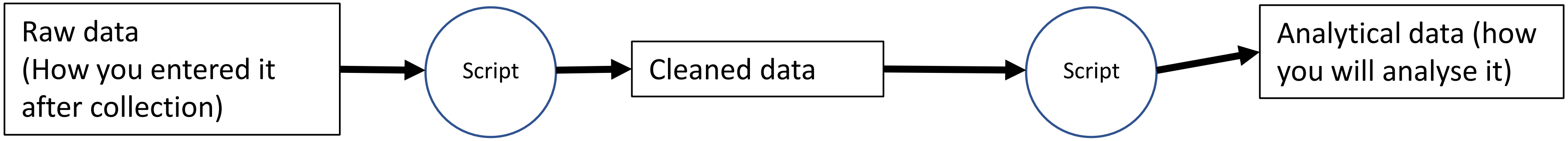


Van Klink


RELATIONSHIP TOOLS
Access
FILE
HOME
CREATE
EXTERNAL DATA
DATABASE TOOLS
DESIGN
Van Klink, Roel (EE)
Edit Relationships
Clear Layout
Relationship Report
Tools
Show Table
Hide Table
Direct Relationships
All Relationships
Close
Relationships
All Access O...
Tables
BuildVersionBugsdata
INDEX
TAttributes
TBiblio
TBiology
TbirmBEETLEdat
TCountry
TCountsheet
TDatesCalendar
TDatesMethods
TDatesPeriod
TDatesRadio
TDatesTypes
TDistrib
TEcoBugs
TEcoDefBugs
TEcoDefGroups
TEcoDefKoch
TEcoKoch
TFossil
TFossilUncertainty
TINDEXReplacements
TKeys
TLab
TLookupCountsheet...
TLookupCountsheet...
TLookupMonths
TLookupUncertainty...
TMCRNames
TMCRSummaryData
TPeriods
TRDB
TRDBCodes
TRDBSystems
TDatesMethods
Abbrev
Method
Type
SortOrder
TDatesPeriod
PeriodDate
SampleCOI
Uncertainty
PeriodCOD
DatingMet
TPeriods
PeriodCODE
PeriodName
PeriodType
PeriodDesc
PeriodRef
PeriodGeog
Begin
BeginBCAD
End
EndBCAD
TBiblio
REFERENCE
AUTHOR
TITLE
Notes
TKeys
CODE
Ref
Data
TAttributes
CODE
AttribType
AttribMeasure
Value
AttribUnits
TRDBSystems
RDBSystemCo
RDBSystem
RDBVersion
RDBSystemDa
RDBFirstPubli
Ref
CountryCode
TTaxoNotes
CODE
Ref
Data
TBiology
CODE
Ref
Data
TMCRNames
MCRName
CompareSt
CODE
tempCODE
MCRNumb
TSiteRef
SiteCODE
Ref
TDistrib
CODE
Ref
Data
TSiteOtherProxies
OtherProx
SiteCODE
HasPollen
HasPlantM
HasDiatom
TSite
SiteCODE
SiteName
Region
Country
NGR
TCountsheet
CountsheetCO
CountsheetNa
SiteCODE
SheetContext
SheetType
TSample
SampleCOD
SiteCODE
X
Y
ZorDepthTo
TFossil
FossilBugsCO
CODE
SampleCODE
Abundance
TFossilUncertai...
FossilBugsCOI
Uncertainty
TEcoBugs
CODE
BugsEcoCODE
TEcoDefBugs
SortOrder
BugsEcoCODE
Definition
Notes
EcoLabel
TSynonym
CODE
SynGenus
SynSpecies
SynAuthority
Ref
Notes
TSynonymNotes
CODE
Notes
TSeasonActiveA...
CODE
HSeason
CountryCode
INDEX
CODE
FAMILY
GENUS
SPECIES
AUTHORITY
TEcoKoch
CODE
BugsKochCod
TDatesCalendar
SampleCODE
Uncertainty
CalendarCODE
Date
BCADBP
DatingMethod
Notes
TCountry
CountryCode
Country
TLookupMonths
SeasonCode
SeasonName
SortOrder
TbirmBEETLEdat
Field1
Field2
Field3
Field4
TEcoDefKoch
KochCode
FullName
KochGroup
Description
Notes
TRDBCodes
RDBCode
Category
RDBDefinition
RDBSystemCo
TRDB
CODE
CountryCode
RDBCode
TSpeciesAssoci...
SpeciesAssoci
CODE
AssociatedSpe
AssociationTy
Ref
TEcoDefGroups
EcoGroupCod
EcoName
TLab
LabID
Labname
Address
Country
Telephone
TDatesRadio
DateCODE
SampleCOD
LabNr
Uncertainty
Date
TMCRSummary...
CODE
TMaxLo
TMaxHi
TMinLo
Ready
NUM LOCK

Van Klink

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


Raw data (How you entered it after collection)

Script

Cleaned data

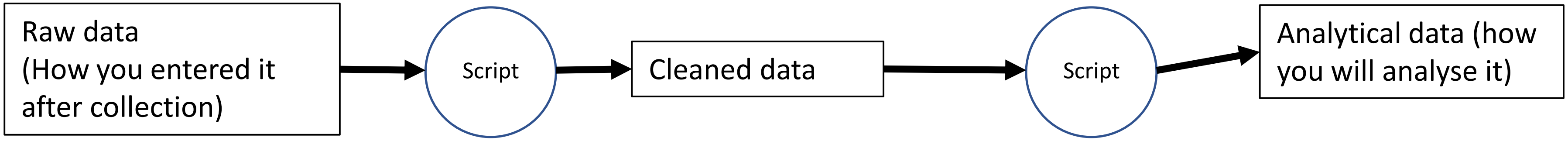
Script

Analytical data (how you will analyse it)

- 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.

Van Klink

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.

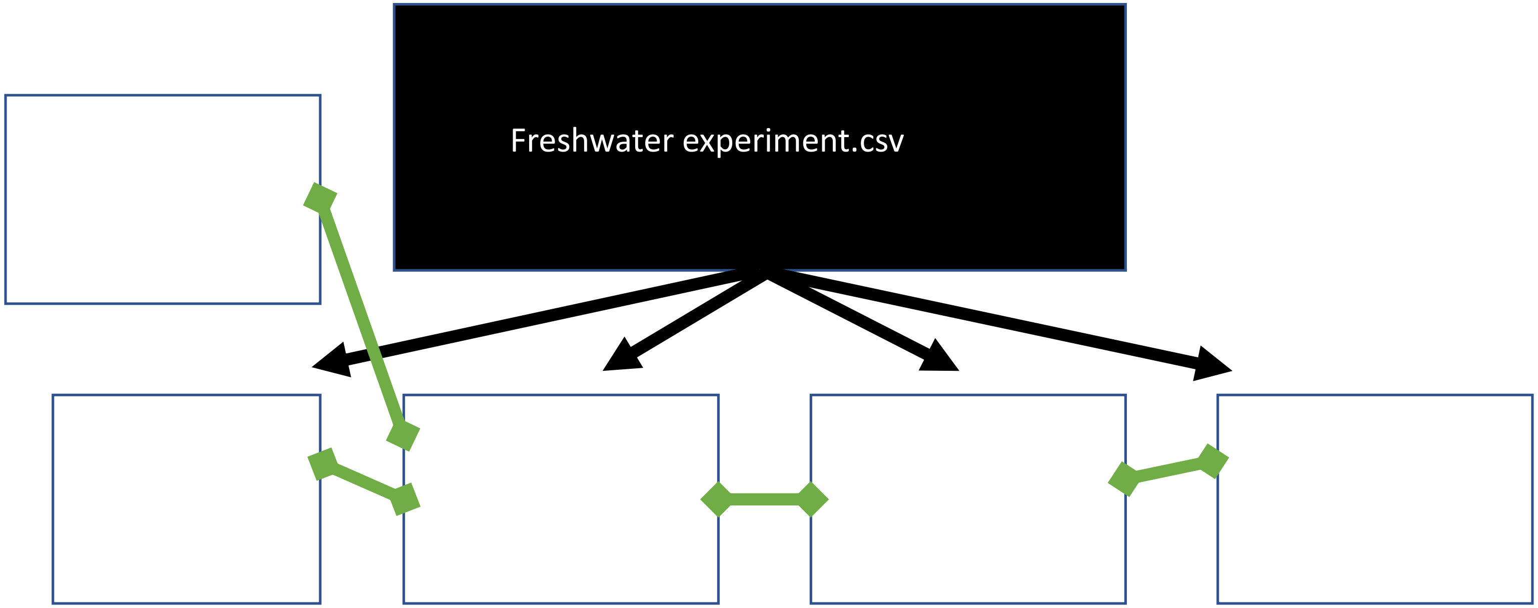
FILE: ‘FRESHWATER EXPERIMENT’

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

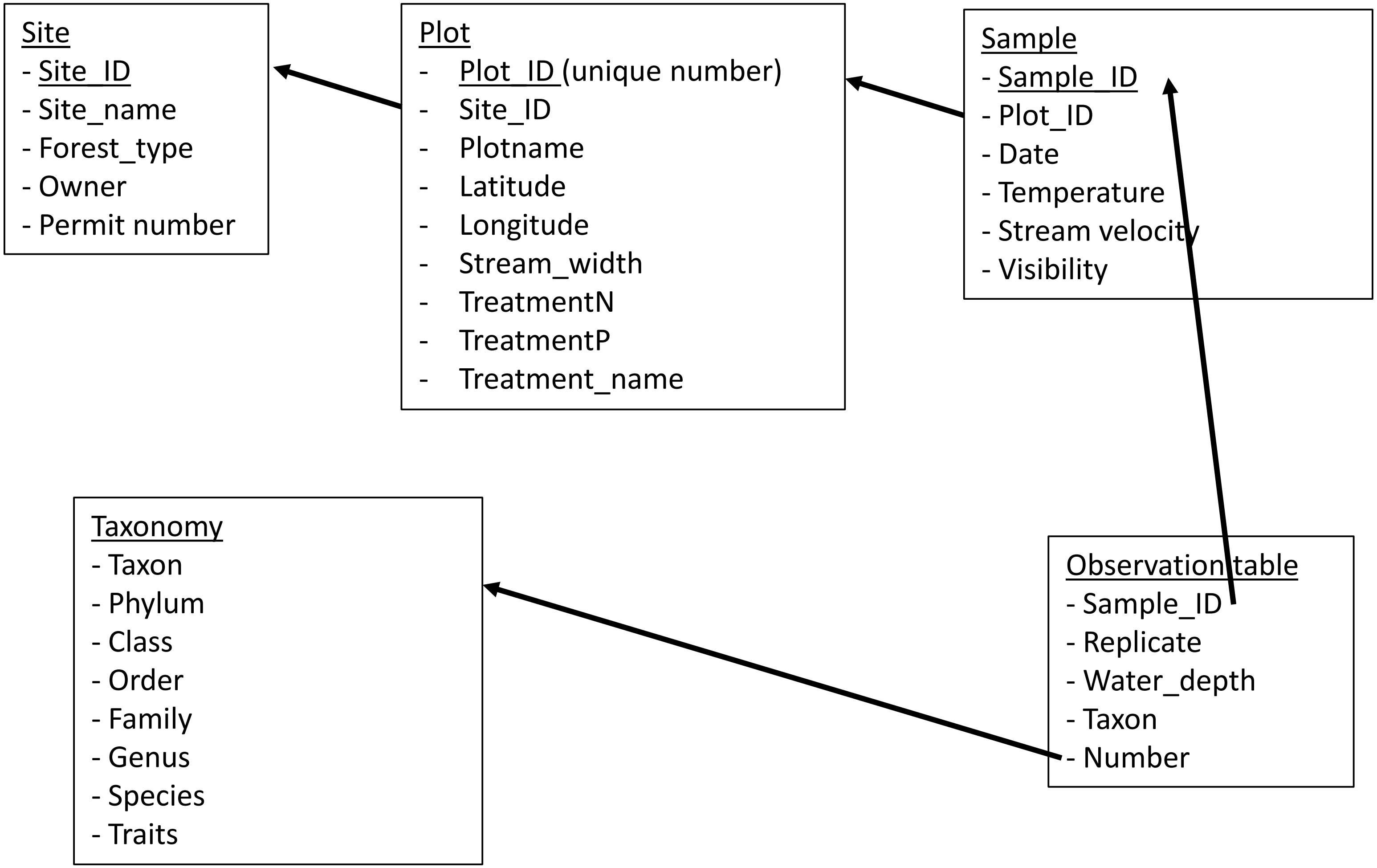
- Brian McGill blogpost + script
- Publications by Borer (2009) and Costello (2014)
- Tidyverse: https://subhayo.wordpress.com/2017/12/16/data-manipulation-of-star-wars-characters-using-dplyr-and-tidyr/
- Pivot tables:

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

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

Van Klink



Site
- Site_ID
- Site_name
- Forest_type
- Owner
- Permit number

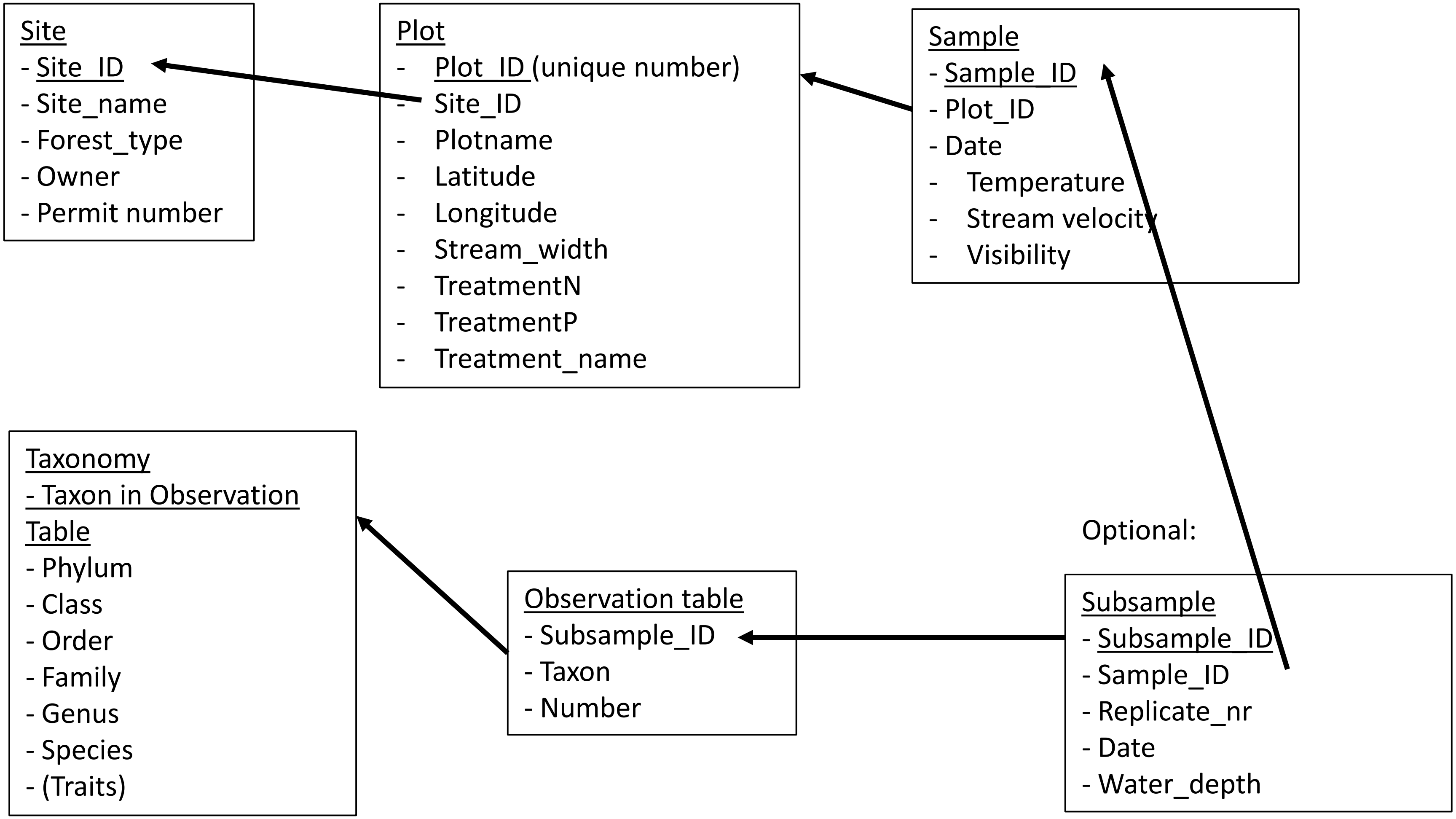
Plot
- Plot_ID (unique number)
- Site_ID
- Plotname
- Latitude
- Longitude
- Stream_width
- TreatmentN
- TreatmentP
- Treatment_name

Sample
- Sample_ID
- Plot_ID
- Date
- Temperature
- Stream velocity
- Visibility

Taxonomy
- Taxon
- Phylum
- Class
- Order
- Family
- Genus
- Species
- Traits

Observation table
- Sample_ID
- Replicate
- Water_depth
- Taxon
- Number

Van Klink



Site
- Site_ID
- Site_name
- Forest_type
- Owner
- Permit number

Plot
- Plot_ID (unique number)
- Site_ID
- Plotname
- Latitude
- Longitude
- Stream_width
- TreatmentN
- TreatmentP
- Treatment_name

Sample
- Sample_ID
- Plot_ID
- Date
- Temperature
- Stream velocity
- Visibility

Taxonomy
- Taxon in Observation Table
- Phylum
- Class
- Order
- Family
- Genus
- Species
- (Traits)

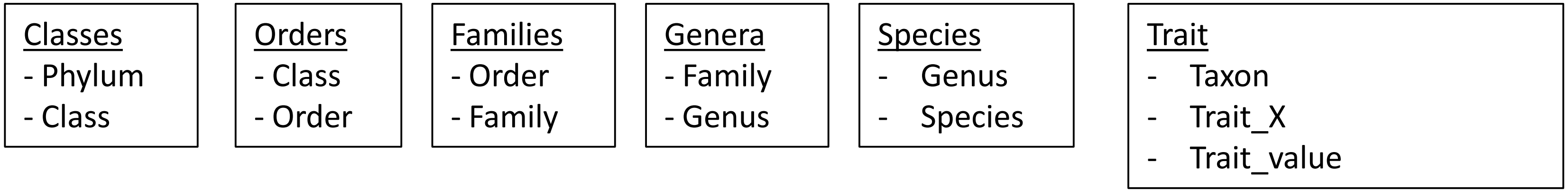
Observation table
- Subsample_ID
- Taxon
- Number

Optional:

Subsample
- Subsample_ID
- Sample_ID
- Replicate_nr
- Date
- Water_depth

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

Anahita Kazem, Feb 2022

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)

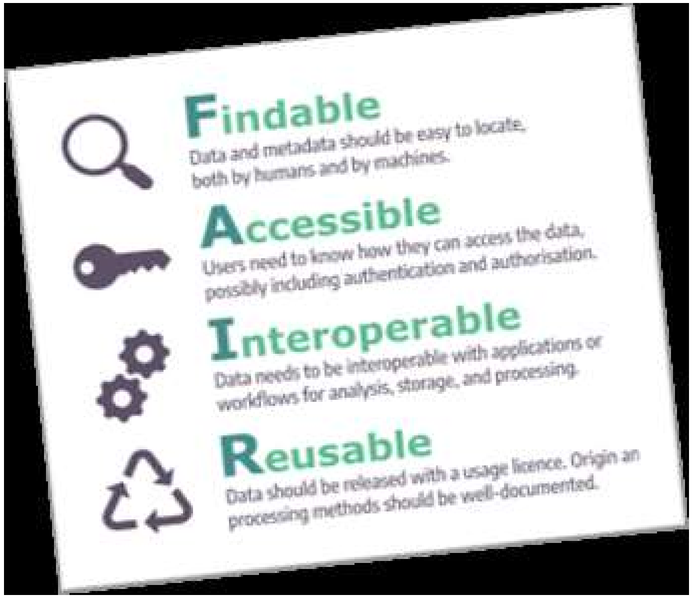
https://opendefinition.org

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 ≥ 1 open source software

Anahita Kazem, Feb 2022

The FAIR data principles



Findable
Data and metadata should be easy to locate, both by humans and by machines.

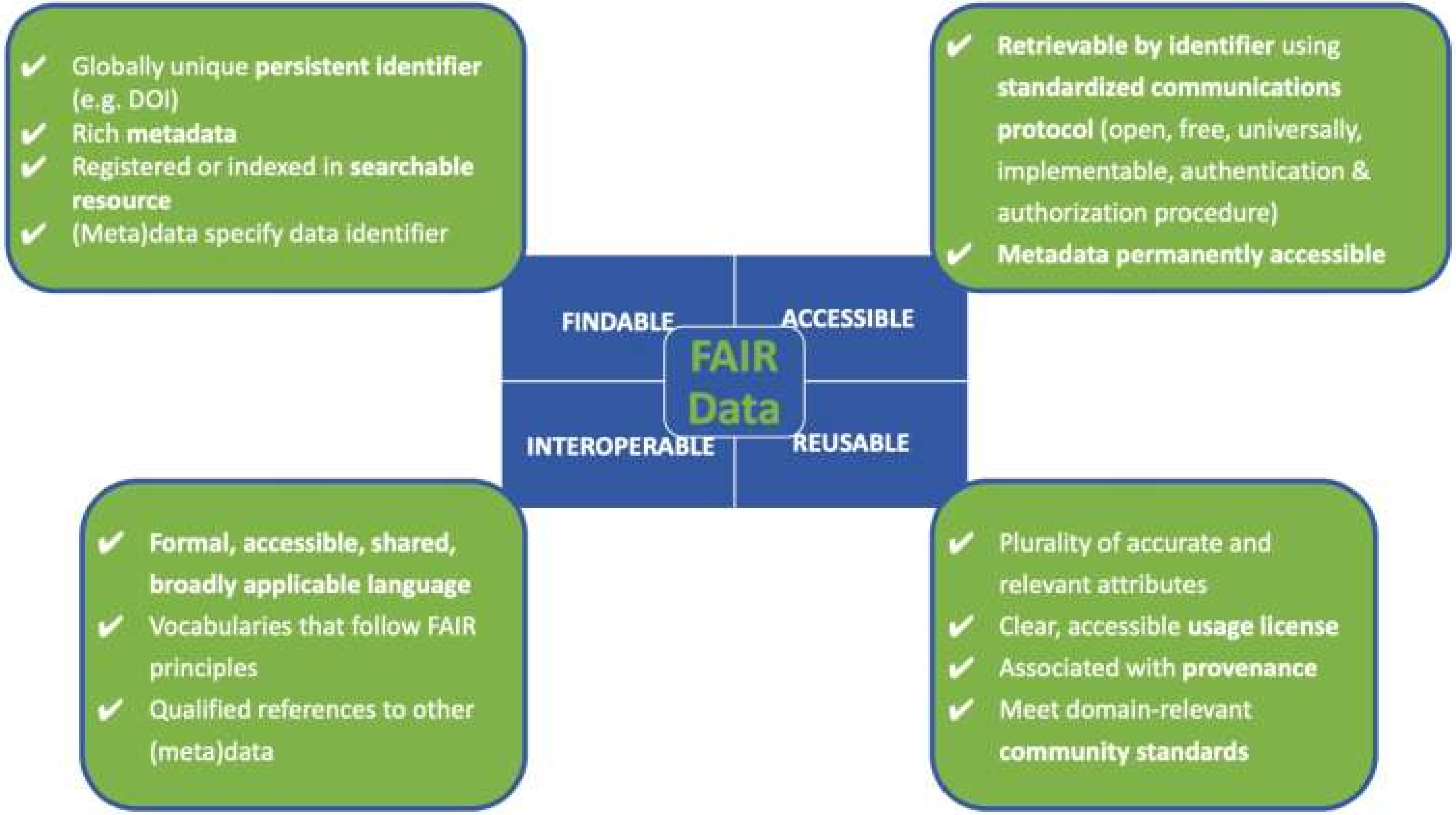
Accessible
Users need to know how they can access the data, possibly including authentication and authorisation.

Interoperable
Data needs to be interoperable with applications or workflows for analysis, storage, and processing.

Reusable
Data should be released with a usage licence. Origin an processing methods should be well-documented.

- 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 ≠ Open - we should strive for FAIR/O



FAIR Data

FINDABLE

- ✔ Globally unique **persistent identifier** (e.g. DOI)
- ✔ Rich **metadata**
- ✔ Registered or indexed in **searchable resource**
- ✔ (Meta)data specify data identifier

ACCESSIBLE

- ✔ **Retrievable by identifier** using **standardized communications protocol** (open, free, universally, implementable, authentication & authorization procedure)
- ✔ **Metadata permanently accessible**

INTEROPERABLE

- ✔ **Formal, accessible, shared, broadly applicable language**
- ✔ Vocabularies that follow FAIR principles
- ✔ Qualified references to other (meta)data

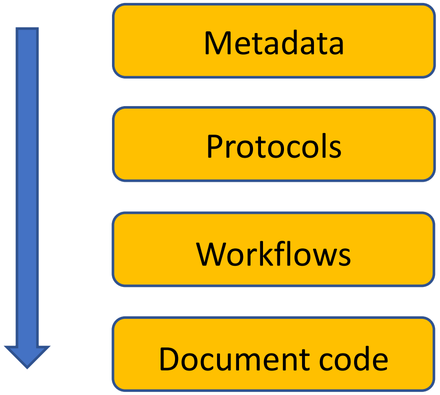
REUSABLE

- ✔ Plurality of accurate and relevant attributes
- ✔ Clear, accessible **usage license**
- ✔ Associated with **provenance**
- ✔ Meet domain-relevant **community standards**

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

Anahita Kazem, Feb 2022

Documenting your work

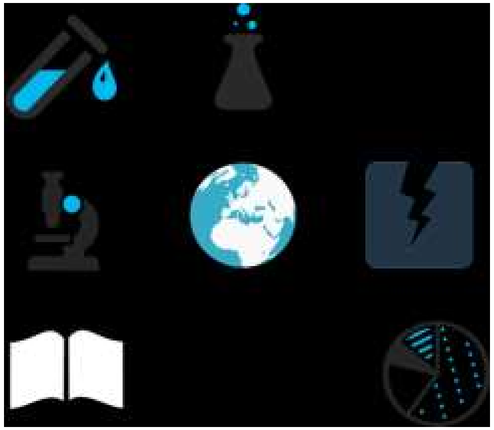


Metadata	Contextualise your data
Protocols	How lab & field work conducted
Workflows	Pre-processing, statistical analysis
Document code	Document the code itself

Do this from the start, and continue doing it throughout

Anahita Kazem, Feb 2022

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

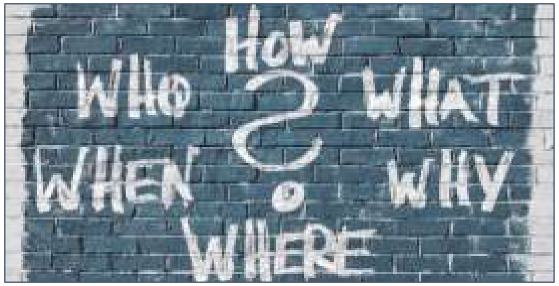
Anahita Kazem, Feb 2022

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:



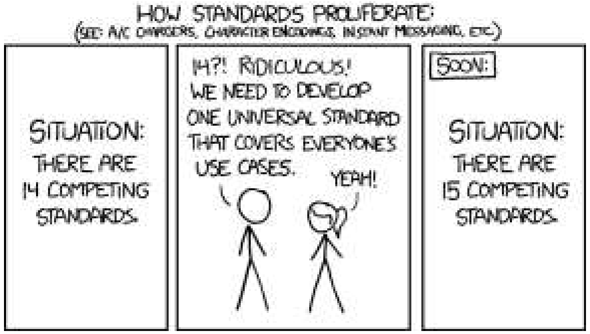

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Image: XKCD, CC BY-NC

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

Anahita Kazem, Feb 2022

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 ≥10 years
- Choose a license

GitHub is not an 'archive'

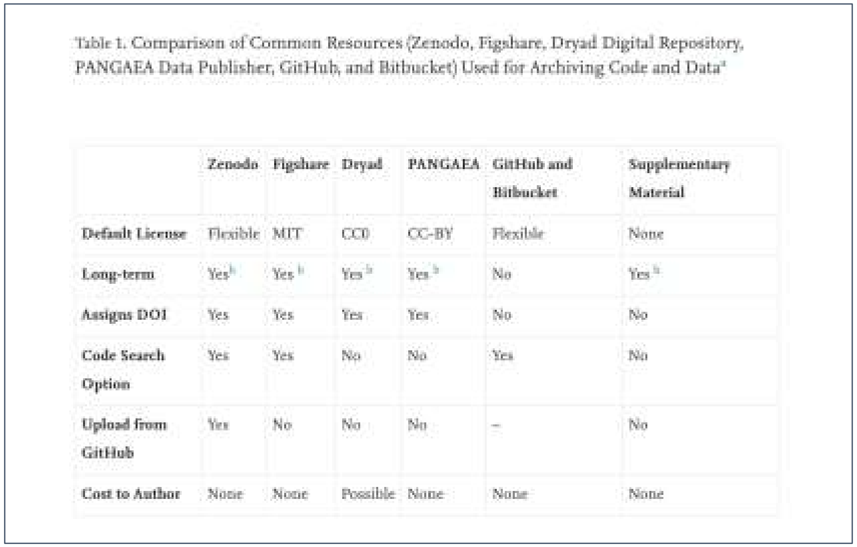


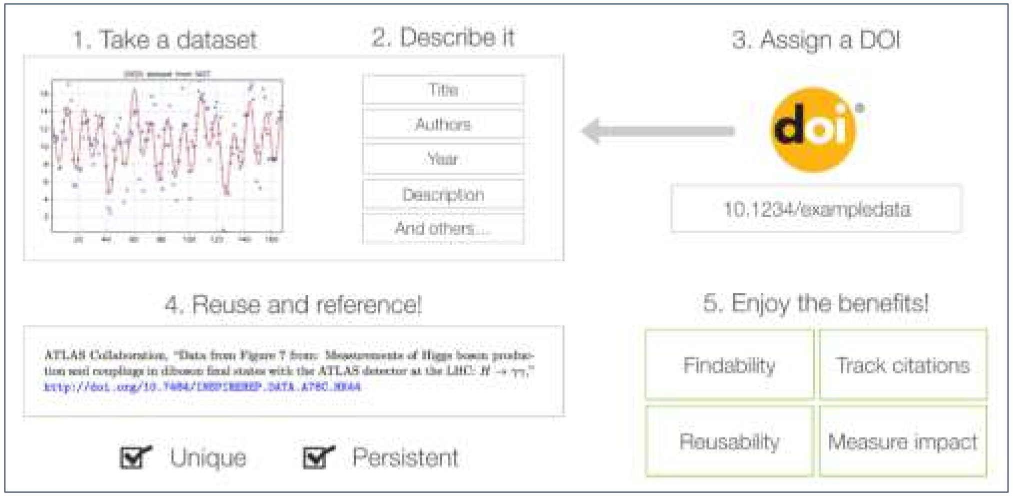
Table 1. Comparison of Common Resources (Zenodo, Figshare, Dryad Digital Repository, PANGAEA Data Publisher, GitHub, and Bitbucket) Used for Archiving Code and Data[a]

	Zenodo	Figshare	Dryad	PANGAEA	GitHub and Bitbucket	Supplementary Material
Default License	Flexible	MIT	CC0	CC-BY	Flexible	None
Long-term	Yes[b]	Yes[b]	Yes[b]	Yes[b]	No	Yes[b]
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



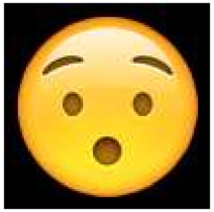
ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \to \gamma\gamma$," http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

doi.datacite.org

Anahita Kazem, Feb 2022

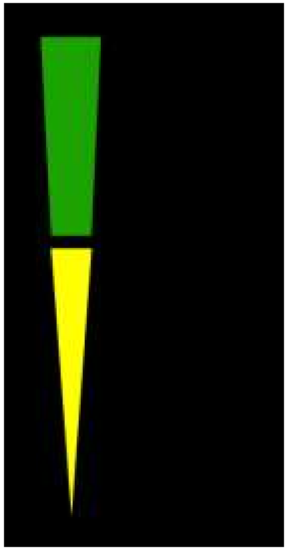
If a dataset/code does not specify a license, then noone can reuse it!

... even if you made the resource easily accessible online



Specify a licence

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



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.

www.creativecommons.org

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 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

Anahita Kazem, Feb 2022

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 Großpösna (51º15'41"N, 12º29'55"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 Groflpˆsna (51∞15'41"N, 12∞29'55"E), with the following number of leaves per tree species: C. betulus ñ 45 leaves, Q. robur ñ 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 Gro▀p÷sna (51░15'41"N, 12░29'55"E), with the following number of leaves per tree species: C. betulus û 45 leaves, Q. robur û 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

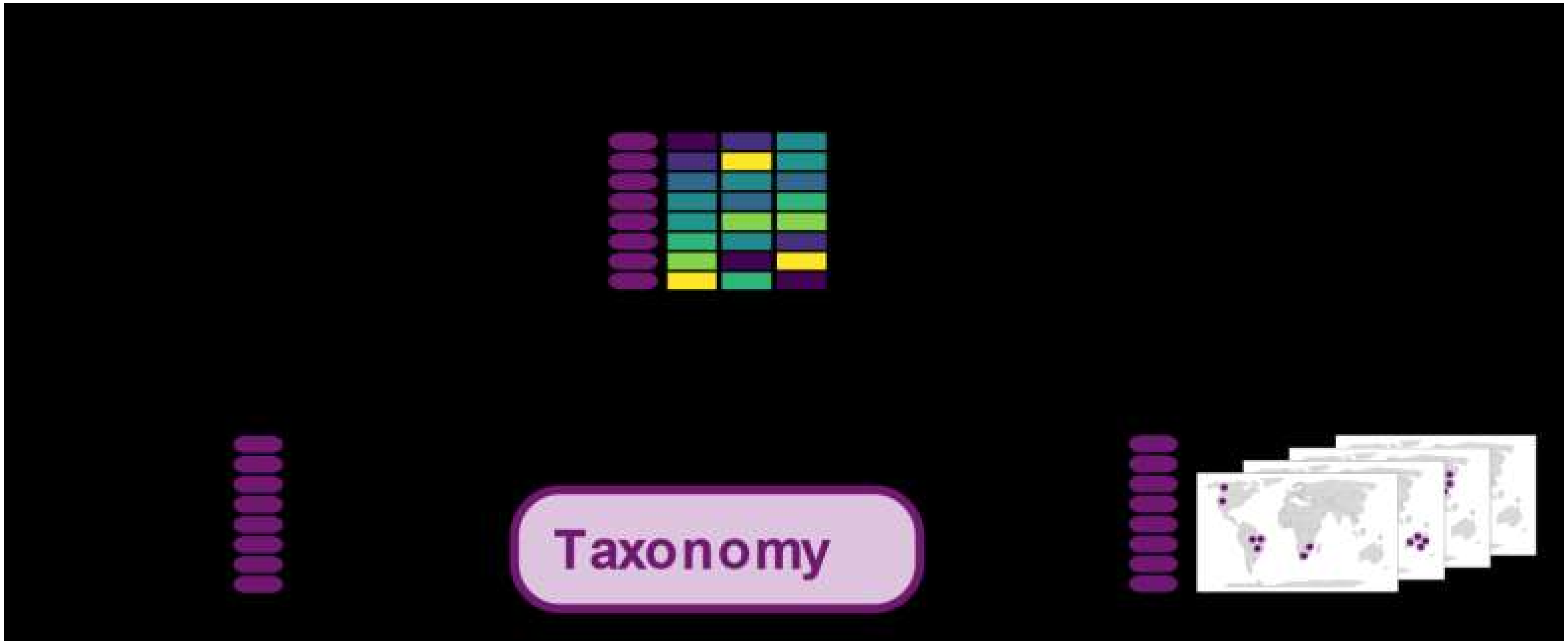
Matching Species Names Across Biodiversity Databases:

Sources, tools, pitfalls and best practices for taxonomic harmonization

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

Berti, Sagouis

Berti, Sagouis



Berti, Sagouis

Dataset A	Dataset B
Sp1	—
Sp2	Sp2
Sp3	Sp3
Sp4	Sp4
—	—
—	Sp6

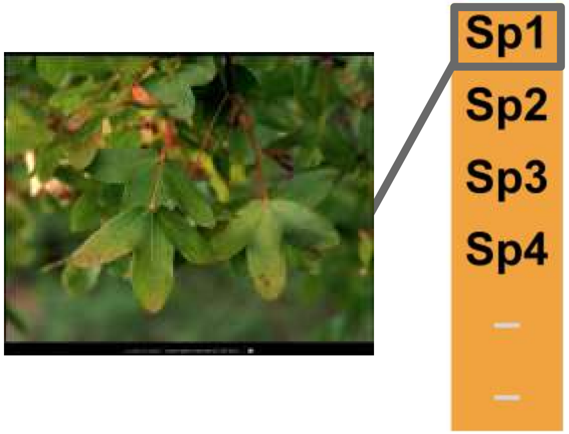
Berti, Sagouis

Dataset A	Dataset B
Sp1	–
Sp2	Sp2
Sp3	Sp3
Sp4	Sp4
–	–
–	Sp6

Acer monspessolanum

Berti, Sagouis

Dataset A



Sp1

Sp2

Sp3

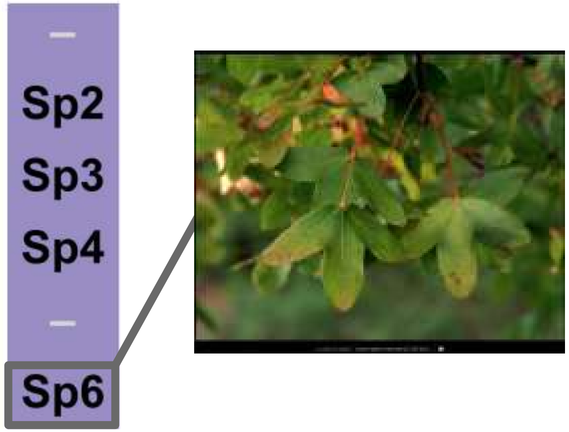
Sp4

–

–

Acer monspessolanum

Dataset B



–

Sp2

Sp3

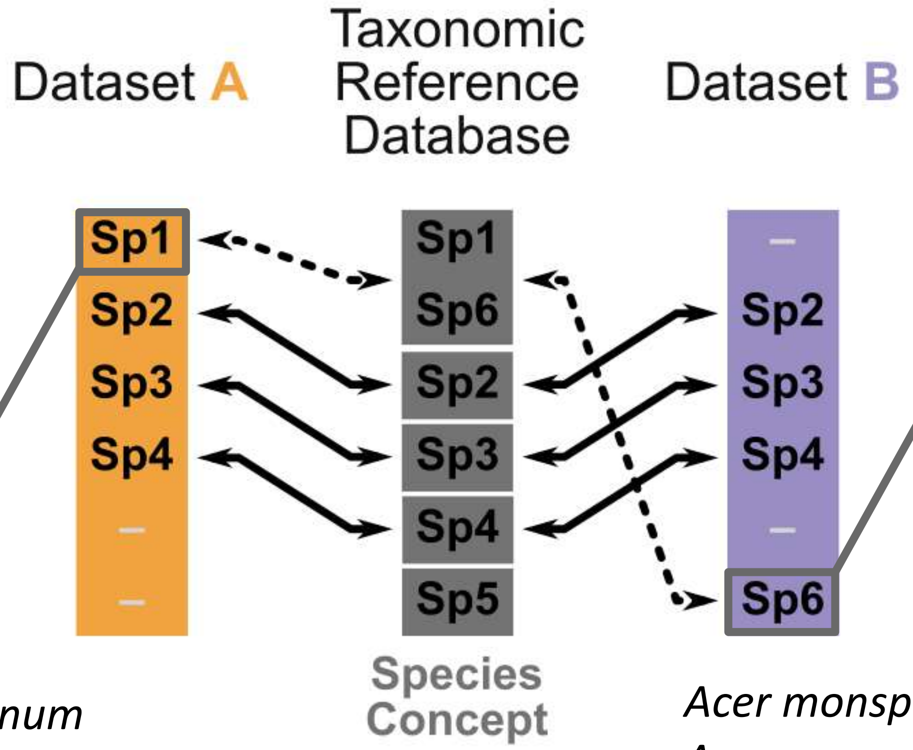
Sp4

–

Sp6

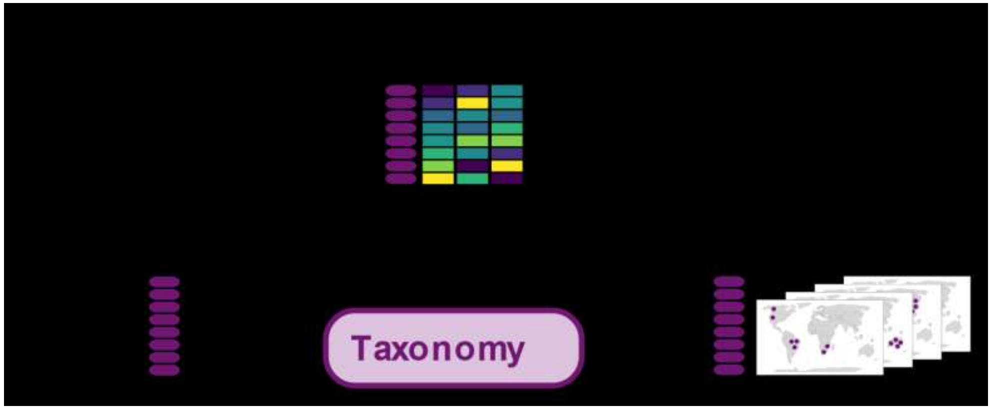
Acer monspessulanum L.
Acer monspessulanum loscosii
Acer monspessolano
Montpellier maple

Berti, Sagouis



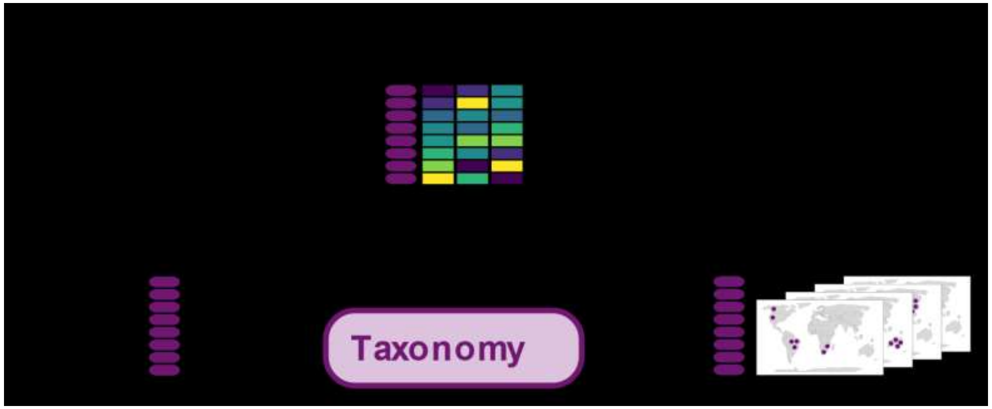
Acer monspessolanum

Acer monspessulanum L.
Acer monspessulanum loscosii
Acer monspessolano
Montpellier maple



Berti, Sagouis

Is there one, best way to harmonize taxonomy?



Berti, Sagouis

Is there one, best way to harmonize taxonomy?

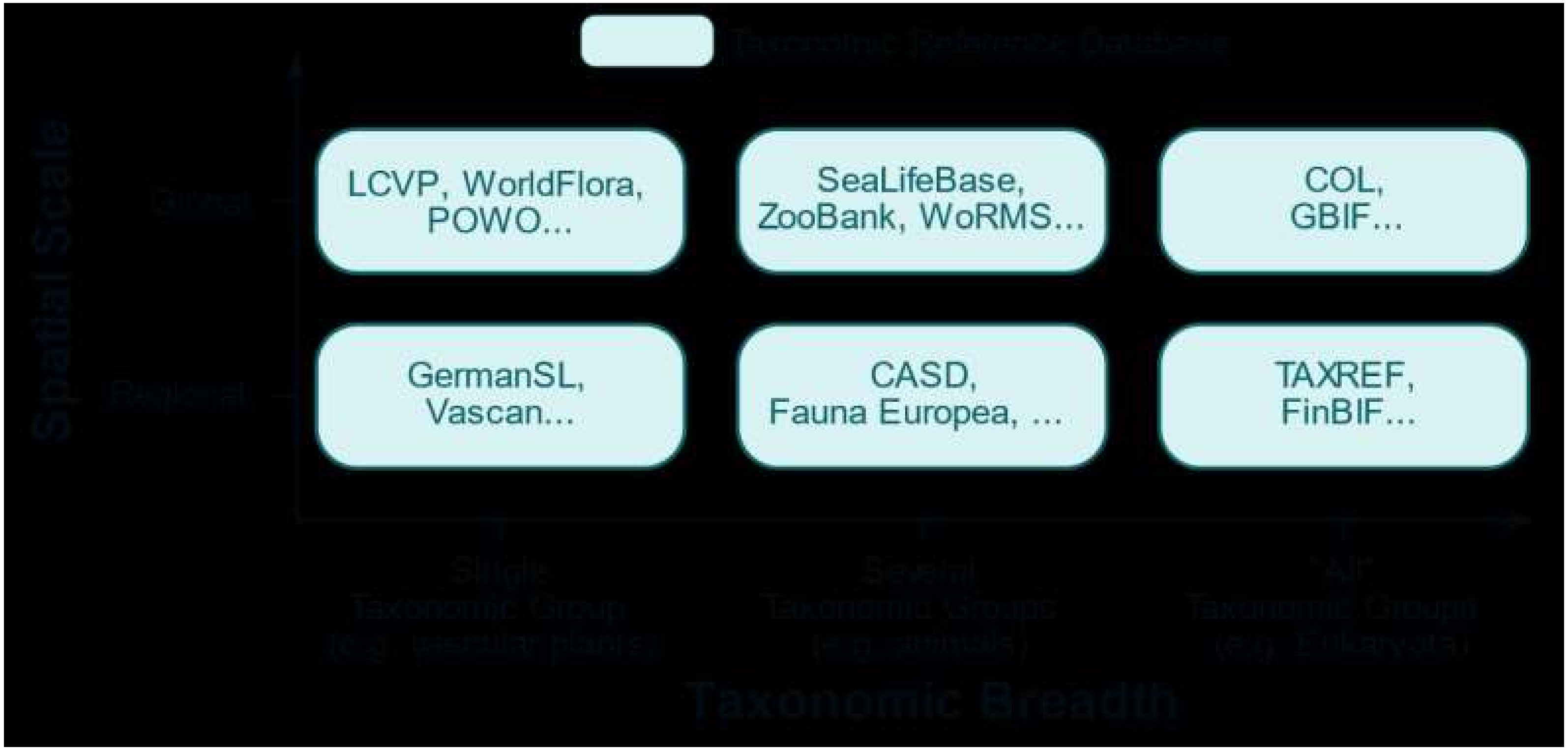
What are the available resources?

- Databases
- Tools (R packages)

Berti, Sagouis

45 databases

A typology of taxonomic databases



LCVP, WorldFlora, POWO...

SeaLifeBase, ZooBank, WoRMS...

COL, GBIF...

GermanSL, Vascan...

CASD, Fauna Europea, ...

TAXREF, FinBIF...

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, **rfishbase**, **rgbif**, ritis, Rocc, rotl, rredlist, rtaxref, SP2000, **taxize**, taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twn, wikitaxa, worms, worrms, zbank, kewr	35

Berti, Sagouis

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, **rfishbase**, **rgbif**, ritis, Rocc, rotl, rredlist, rtaxref, SP2000, **taxize**, taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twn, wikitaxa, worms, worrms, zbank, kewr	35
Database access (offline)	**lcvplants**, ncbit, splister, taxadb, taxalight, taxastand, taxizedb, taxonlookup, **vegdata**, WorldFlora	10

Berti, Sagouis

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, **rfishbase**, **rgbif**, ritis, Rocc, rotl, rredlist, rtaxref, SP2000, **taxize**, taxonomizr, taxonomyCleanr, Taxonstand, taxotools, taxreturn, TNRS, tpl, twn, wikitaxa, worms, worrms, zbank, kewr	35
Database access (offline)	**lcvplants**, ncbit, splister, taxadb, taxalight, taxastand, taxizedb, taxonlookup, **vegdata**, WorldFlora	10
Data wrangling	metacoder, monographR, **rgnparser**, splister, taxastand, taxreturn, taxspell, traitdatafrom, vegdata, vegtable, yatah	11

61 packages

Available R packages

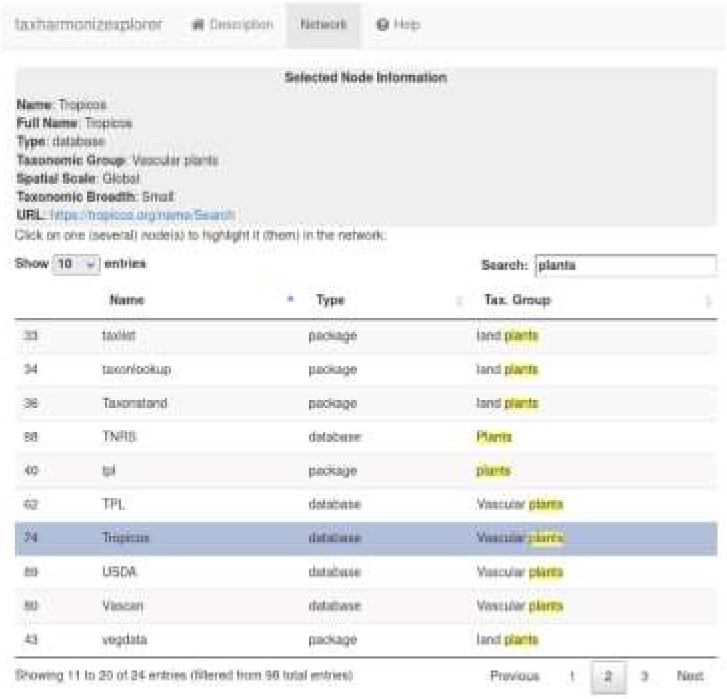
Category	Packages	
Infrastructure	taxa, taxlist, taxview	3
Database access (online)	algaeClassify, AmphiNom, dyntaxa, finbif, flora, mammals, natserv, neotoma2, paleobioDB, plantlist, rcol, rebird, rentrez, **rfishbase**, **rgbif**, ritis, Rocc, rotl, rredlist, rtaxref, SP2000, **taxize**, taxonomizr,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

Berti, Sagouis

45 Databases and relationships among them

35 R packages accessing databases

61 R packages and relationships between them



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

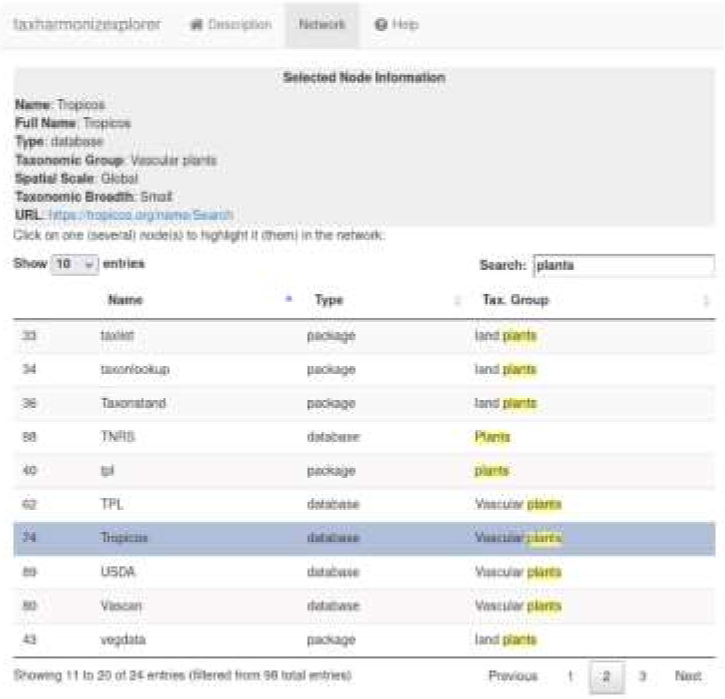
Search: plants

	Name	Type	Tax. Group
33	taxlist	package	land plants
34	taxonlookup	package	land plants
36	Taxonstand	package	land plants
58	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

Berti, Sagouis



taxharmonizexplorer | 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.

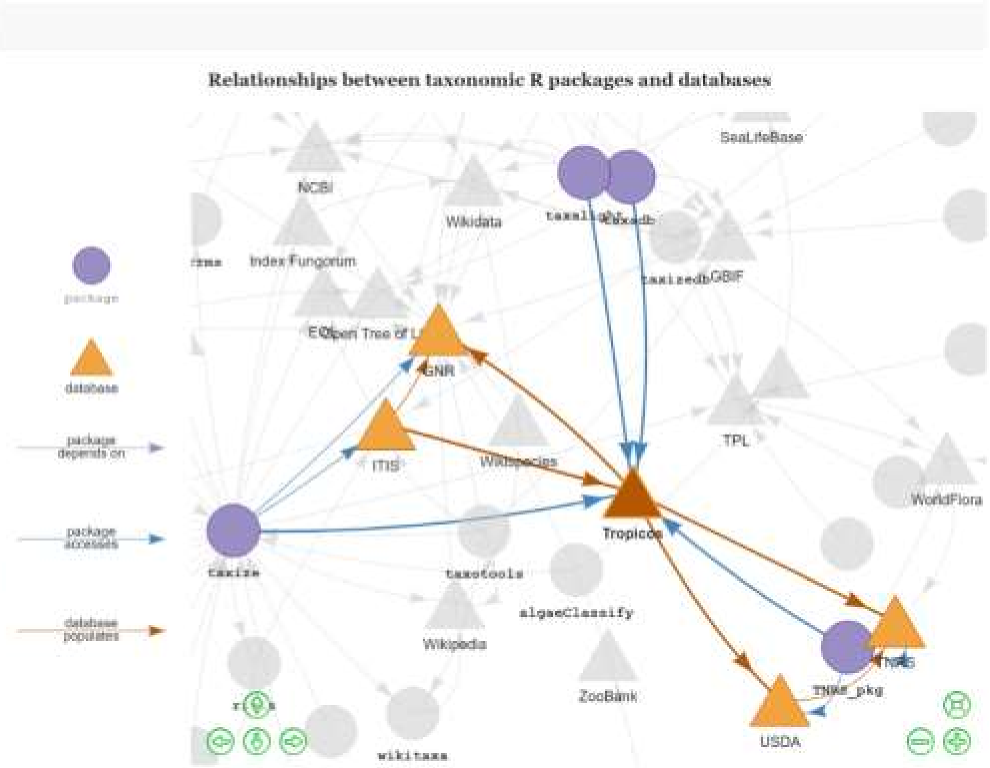
Show 10 entries

Search: plants

	Name	Type	Tax. Group
33	taxlist	package	land plants
34	taxonlookup	package	land plants
36	Taxonstand	package	land plants
58	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

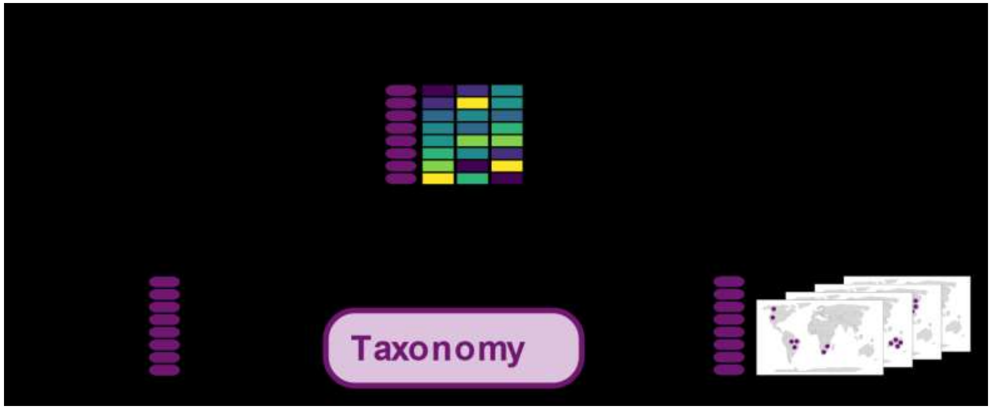
Showing 11 to 20 of 24 entries (filtered from 98 total entries)

Previous 1 2 3 Next



Relationships between taxonomic R packages and databases

Berti, Sagouis



Is there one, best way to harmonize taxonomy?

What are the available resources?

- Databases
- Tools (R packages)

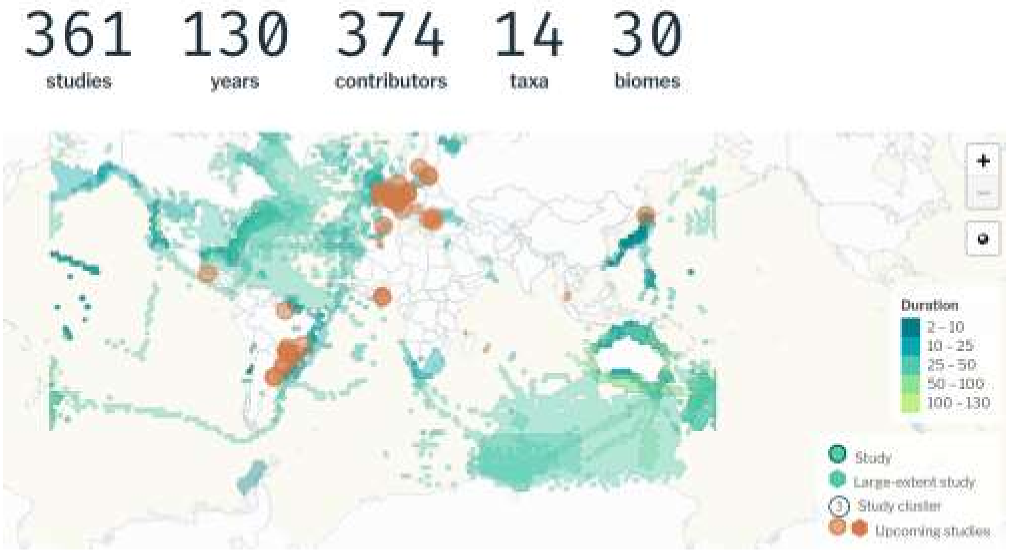
Which workflows are most suitable?

- Accurate
- Easy to implement
- Matching most names

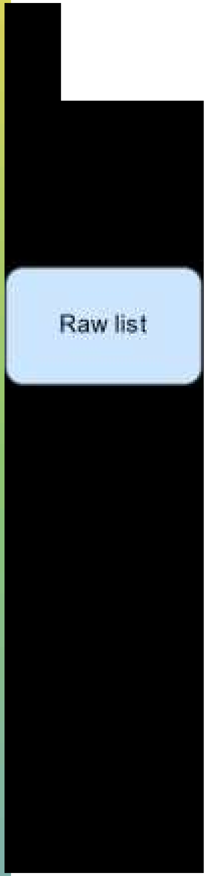
Berti, Sagouis

Harmonize BioTIME database

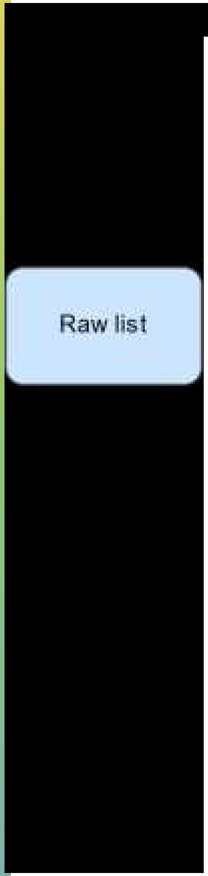
Global database of assemblage time series



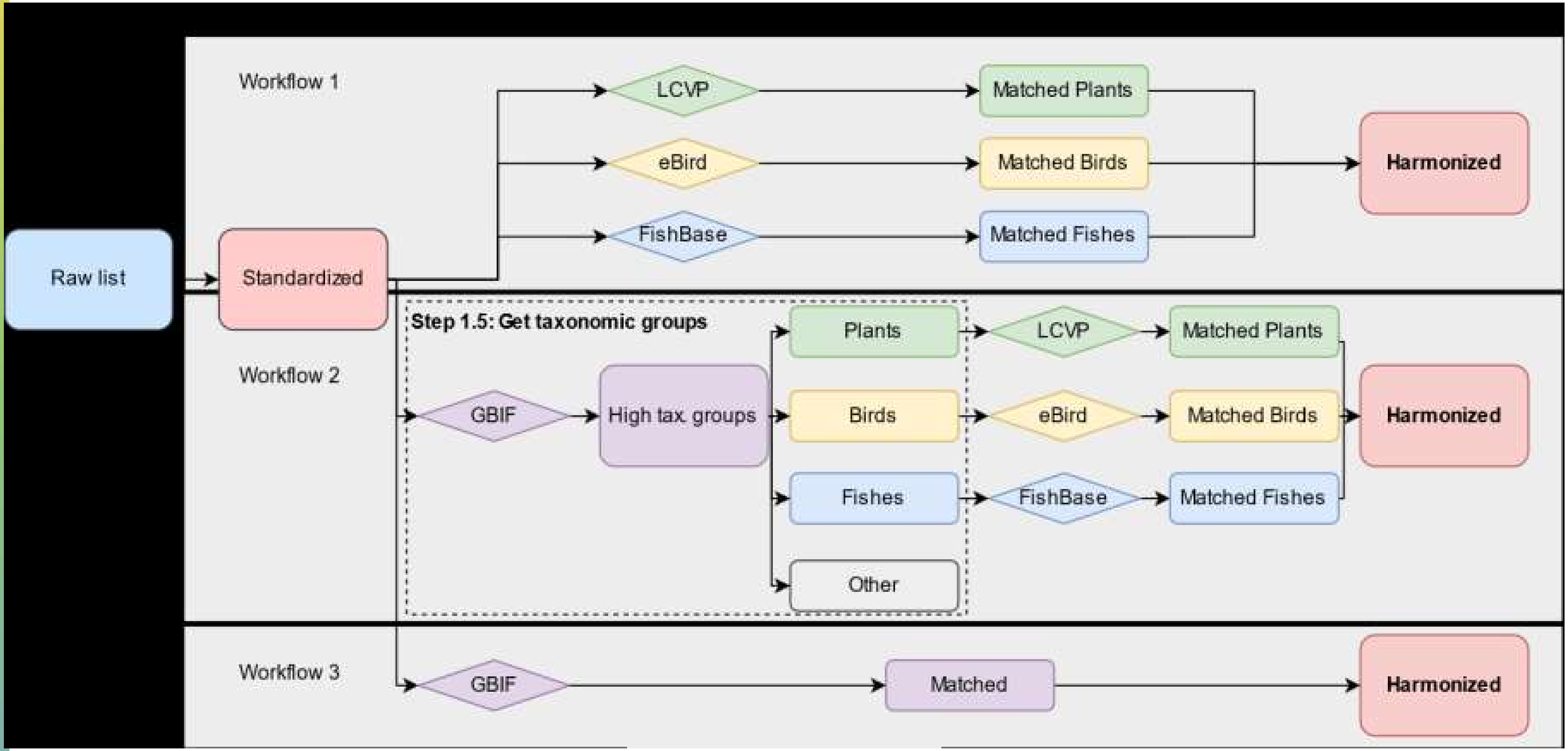
Only birds, fishes, and vascular plants



Raw list



gn_parse_tidy() from package **rgnparser** v.0.2.0


Workflow 1
LCVP
Matched Plants
eBird
Matched Birds
Harmonized
FishBase
Matched Fishes
Raw list
Standardized
Step 1.5: Get taxonomic groups
Plants
LCVP
Matched Plants
Workflow 2
GBIF
High tax. groups
Birds
eBird
Matched Birds
Harmonized
Fishes
FishBase
Matched Fishes
Other
Workflow 3
GBIF
Matched
Harmonized

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-selecter/

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

Thank you