#BOSC2014

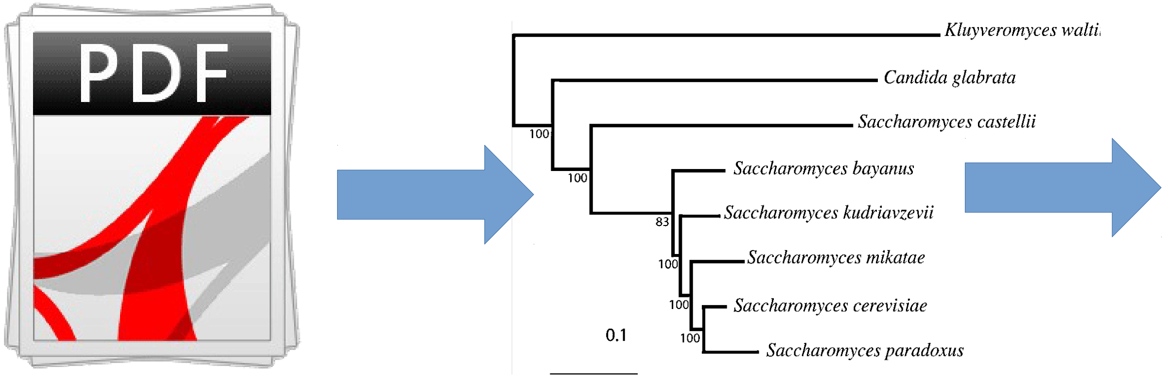
PLUTo

Phyloinformatic

Literature

Unlocking

Tools



Machine-Readable

Re-usable

Data & Metadata

Ross Mounce @rmounce
Peter Murray-Rust @petermurrayrust
Matthew Wills

BBSRC
bioscience for the future

UNIVERSITY OF
BATH

Talk Structure

- **Why re-extract?**
  Everyone shares their data, right? [no]

- **Where are the trees?**
  Creating an atlas of phylogeny

- **How to scalably extract tree data?**
  Liberating Figure Images & Captions
  Extracting Re-usable Data from Images

These slides are also up on slideshare

Why hack data from the literature?

Multiple independent studies show **re-usable phylogenetic data is NOT publicly available online for most studies**

- Stoltzfus *et al.* (2012) BMC Research Notes estimates 4%
- Drew *et al.* (2013) PLOS Biology estimates 17%
- Magee *et al.* (2014) arXiv preprint, estimates 25%

Why the difference between studies? Different methods & scope
Drew & Magee sampled only from 'better' papers

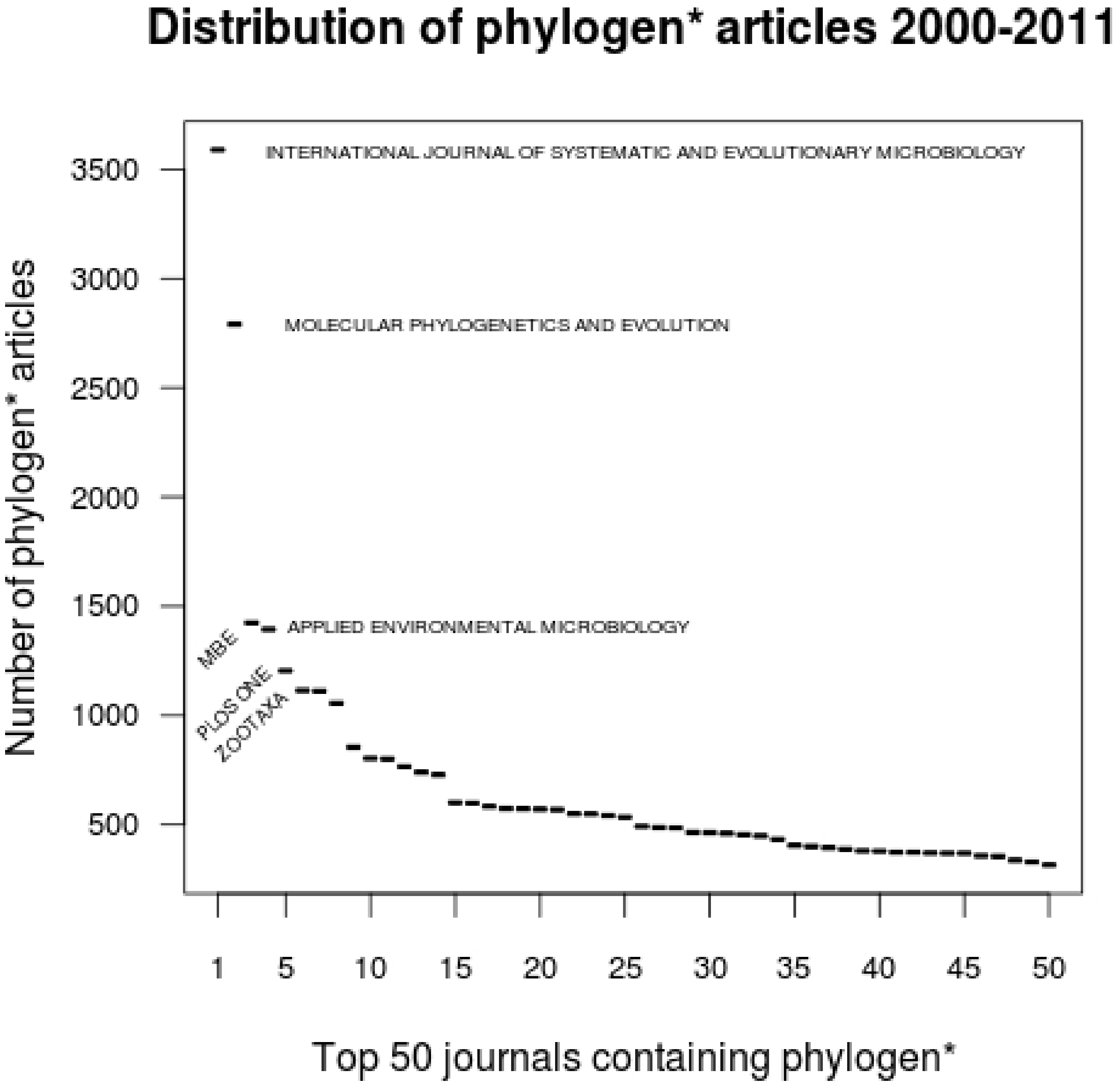
Drew: from well-known journals (only), excluding less-read journals
Magee: from papers citing relatively new, complex methods

Over ALL journals/papers Stoltzfus (2012) probably provides the most representative estimate

Pop Quiz Time

Which journal publishes the most papers containing phylogenetic analyses, per year?

Credit: thanks @rdmpage for the 'pop quiz' technique



#1 is IJSEM

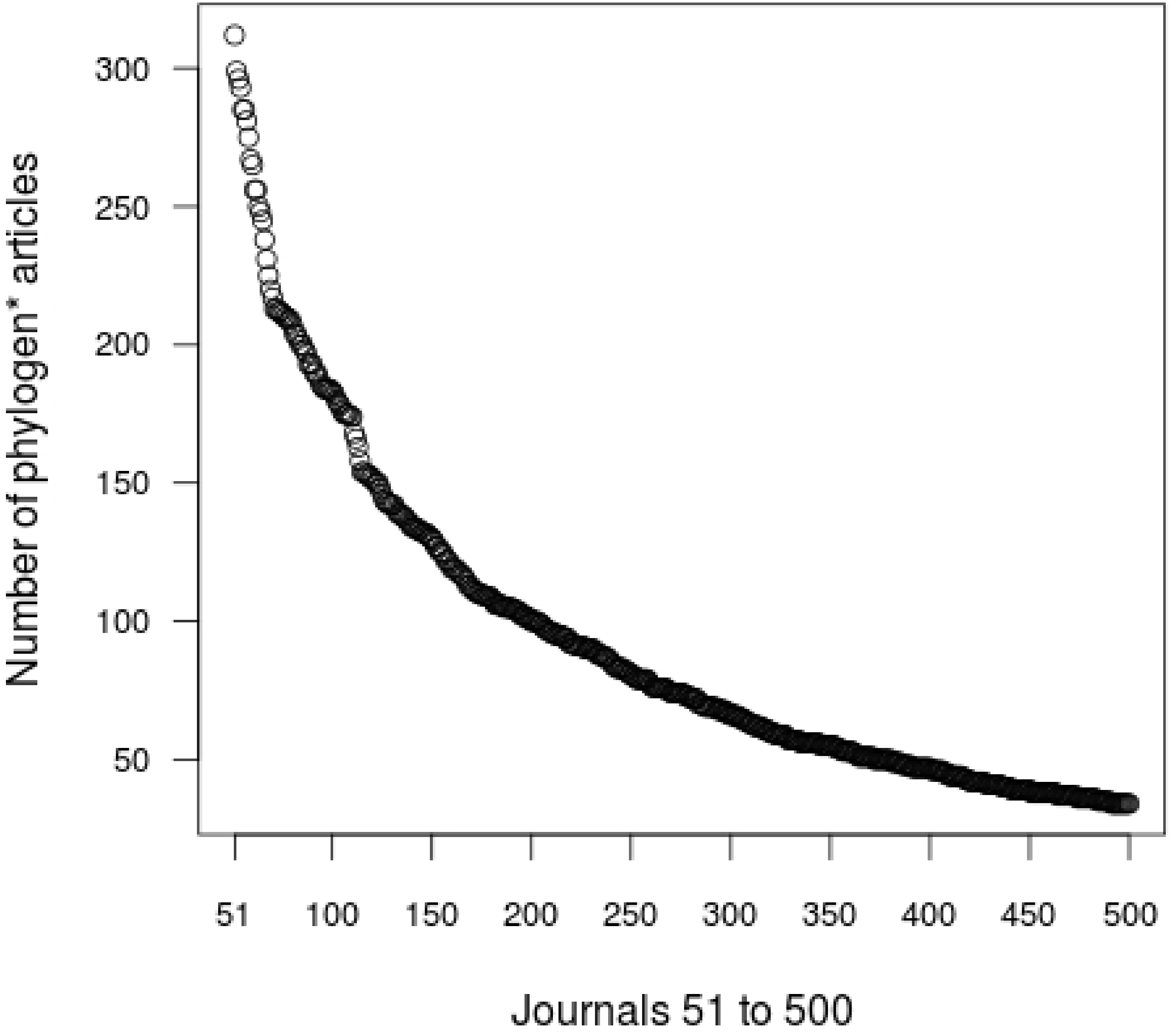
International Journal of Systematic & Evolutionary Microbiology

#5 PLOS ONE

(probably #3 now)

Source: Web of Science /
Mounce (2013) PhD thesis

The long tail distribution of phylogenetic analyses



There's at least a 1000 different journals in which phylogenetic analyses have been published in.

Collectively this represents significant volume.

In terms of journals, volume of phylogeny papers published has no relation to 'quality' of phylogenetic analysis

Creating an atlas of phylogeny

Problems:

- Indexers like Google Scholar, Scopus & Web of Science don't perfectly index the literature – many false negatives (relevant papers not found that should be found)

- No-one has access to ALL journals. Paywalls. Grr

- Even *with* legitimate access, publisher-imposed & copyright restrictions hamper phylogeny discovery

Solutions (partial):

- As of June 1st 2014 the UK has new copyright exceptions to enable and protect text & data mining for non-commercial research purposes [link]

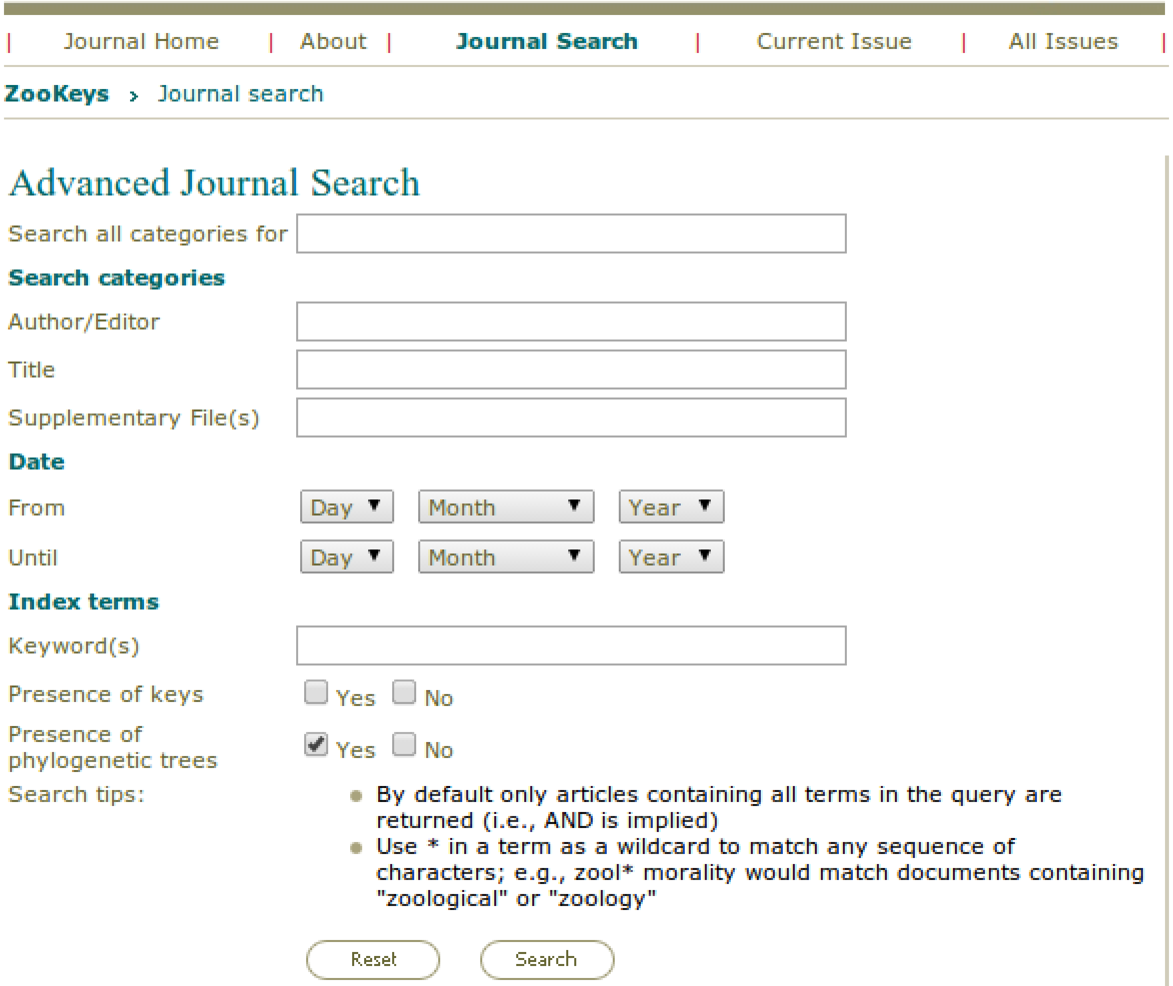
Searching for phylogeny is hard

A peer-reviewed open-access journal

ZooKeys

Launched to accelerate biodiversity research

Make it a lot easier!


| Journal Home | About | Journal Search | Current Issue | All Issues |
ZooKeys › Journal search

Advanced Journal Search

Search all categories for	
Search categories	
Author/Editor	
Title	
Supplementary File(s)	
Date	
From	Day ▼ Month ▼ Year ▼
Until	Day ▼ Month ▼ Year ▼
Index terms	
Keyword(s)	
Presence of keys	☐ Yes ☐ No
Presence of phylogenetic trees	☑ Yes ☐ No

Search tips:

- By default only articles containing all terms in the query are returned (i.e., AND is implied)
- Use * in a term as a wildcard to match any sequence of characters; e.g., zool* morality would match documents containing "zoological" or "zoology"

Search by “presence of phylogenetic trees”

Link to journal search here

Creating an OA atlas of phylogeny

flickr

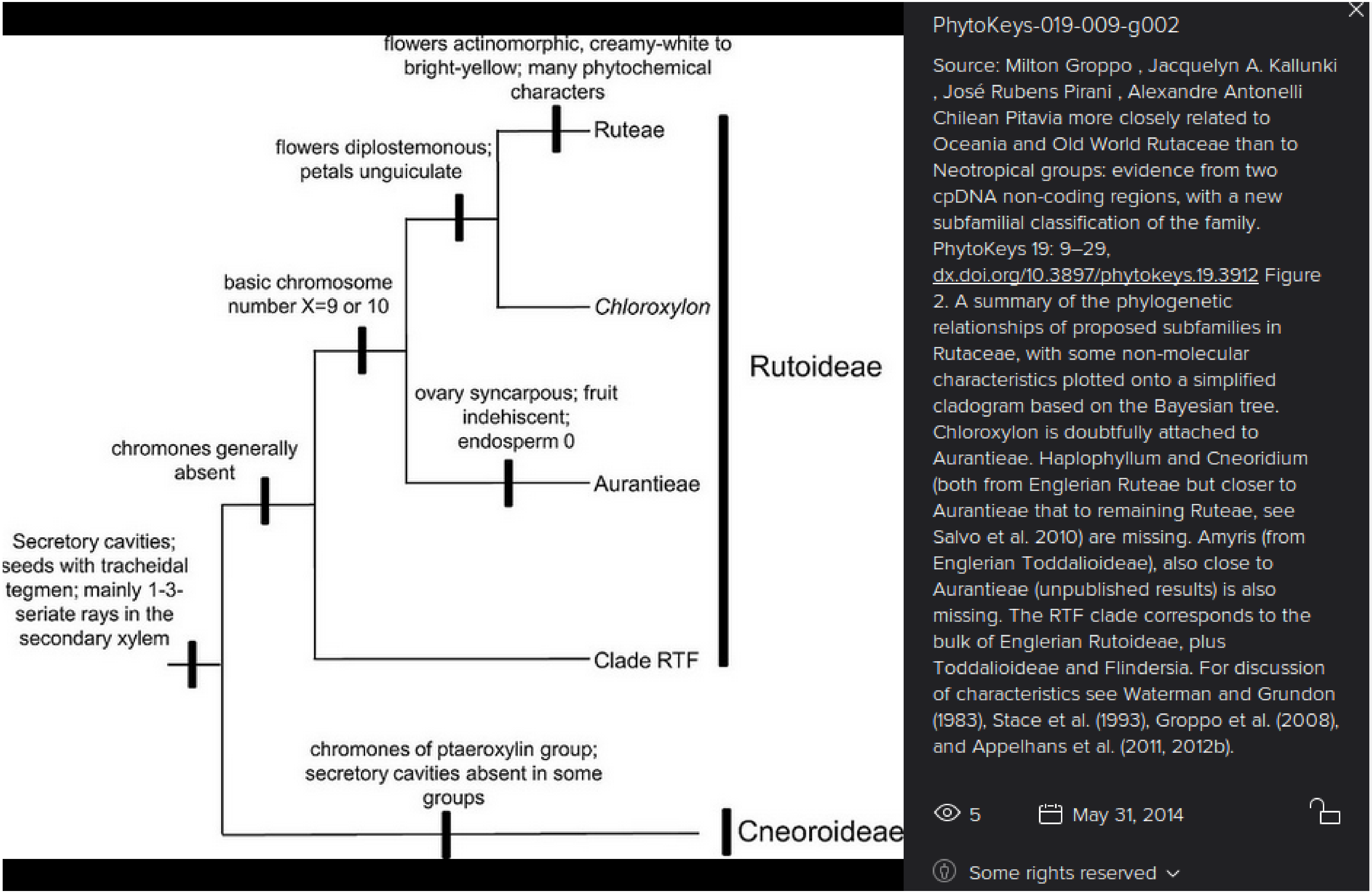
- Free-to-use platform (free as in beer, it's not open)

One Terabyte of free storage per account

- Highly popular platform for image sharing
  (in top 100 most frequently visited websites of the world)

- Supports Creative Commons licensing (many platforms don't)

- Feature-rich, good UI, useful API, etc...

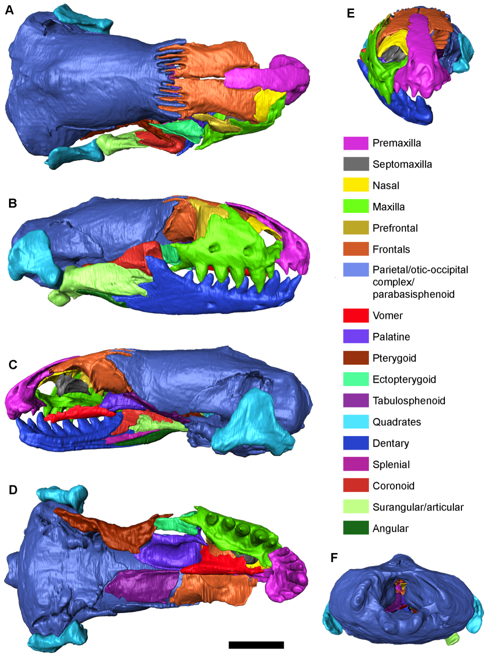

flowers actinomorphic, creamy-white to bright-yellow; many phytochemical characters
Ruteae
flowers diplostemonous; petals unguiculate
basic chromosome number X=9 or 10
Chloroxylon
Rutoideae
ovary syncarpous; fruit indehiscent; endosperm 0
chromones generally absent
Aurantieae
Secretory cavities; seeds with tracheidal tegmen; mainly 1-3-seriate rays in the secondary xylem
Clade RTF
chromones of ptaeroxylin group; secretory cavities absent in some groups
Cneoroideae
5
May 31, 2014
Some rights reserved

PhytoKeys-019-009-g002

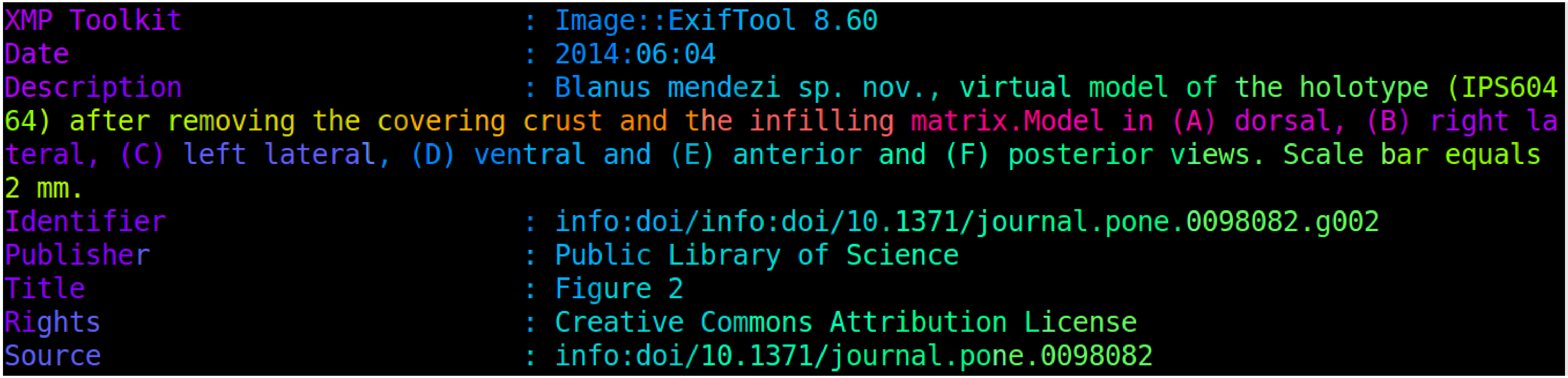
Source: Milton Groppo , Jacquelyn A. Kallunki , José Rubens Pirani , Alexandre Antonelli Chilean Pitavia more closely related to Oceania and Old World Rutaceae than to Neotropical groups: evidence from two cpDNA non-coding regions, with a new subfamilial classification of the family. PhytoKeys 19: 9–29, dx.doi.org/10.3897/phytokeys.19.3912 Figure 2. A summary of the phylogenetic relationships of proposed subfamilies in Rutaceae, with some non-molecular characteristics plotted onto a simplified cladogram based on the Bayesian tree. Chloroxylon is doubtfully attached to Aurantieae. Haplophyllum and Cneoridium (both from Englerian Ruteae but closer to Aurantieae that to remaining Ruteae, see Salvo et al. 2010) are missing. Amyris (from Englerian Toddalioideae), also close to Aurantieae (unpublished results) is also missing. The RTF clade corresponds to the bulk of Englerian Rutoideae, plus Toddalioideae and Flindersia. For discussion of characteristics see Waterman and Grundon (1983), Stace et al. (1993), Groppo et al. (2008), and Appelhans et al. (2011, 2012b).

Full attribution visible next to figure. One-click link to source. Full caption text. Searchable. View-counter (METRICS!). Open licencing marked (tells you it's CC BY on mouse-over)

Only one publisher currently embeds useful metadata in their figure images



Well done PLOS!
Not perfect though.
Author names &
the paper title are
NOT embedded


```
XMP Toolkit                     : Image::ExifTool 8.60
Date                            : 2014:06:04
Description                     : Blanus mendezi sp. nov., virtual model of the holotype (IPS604
64) after removing the covering crust and the infilling matrix.Model in (A) dorsal, (B) right la
teral, (C) left lateral, (D) ventral and (E) anterior and (F) posterior views. Scale bar equals
2 mm.
Identifier                      : info:doi/info:doi/10.1371/journal.pone.0098082.g002
Publisher                       : Public Library of Science
Title                           : Figure 2
Rights                          : Creative Commons Attribution License
Source                          : info:doi/10.1371/journal.pone.0098082
```

The OA 'Atlas of Phylogeny' nearly 10,000 figures!

- 4045 phylogeny figures from PLOS ONE
  - bit.ly/PLOStrees
- 5215 phylogeny figures from 154 OA journals (Pensoft, BMC, FrontiersIn, other PLOS journals, Hindawi, MDPI) & a tiny number of hybrid OA papers from Elsevier, Royal Society and Magnolia Press.
  - bit.ly/phylofigs

correct as of June 22nd 2014

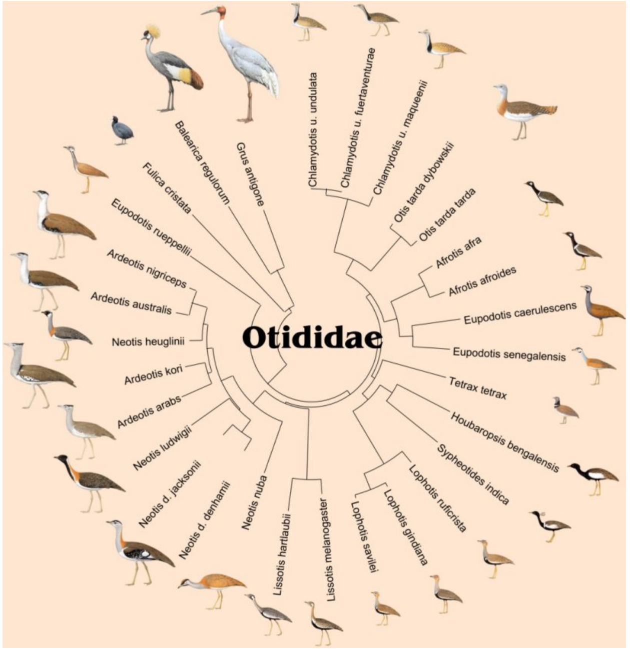
How to get the data from the image?

- Previous work

TreeThief (Rambaut, 2000) old, not used anymore

TreeRipper (Hughes, 2011) automated, but v. picky

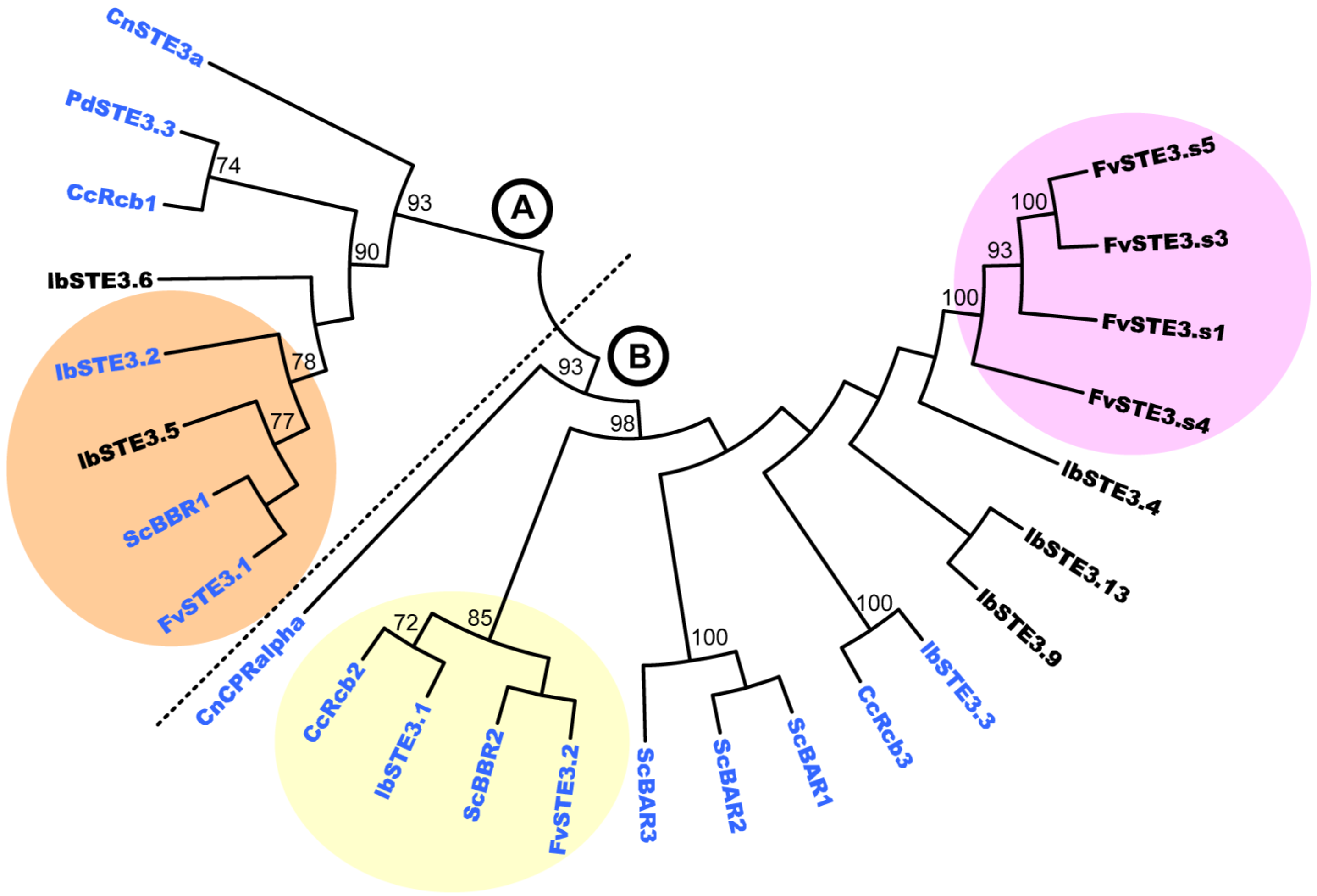
TreeSnatcher Plus (Laubach *et al.* 2012) manual



TreeSnatcher authors report it took them **21 minutes** to manually extract the tree & taxon labels from this radial bustard tree, using TreeSnatcher Plus (Supp. Data. 6)

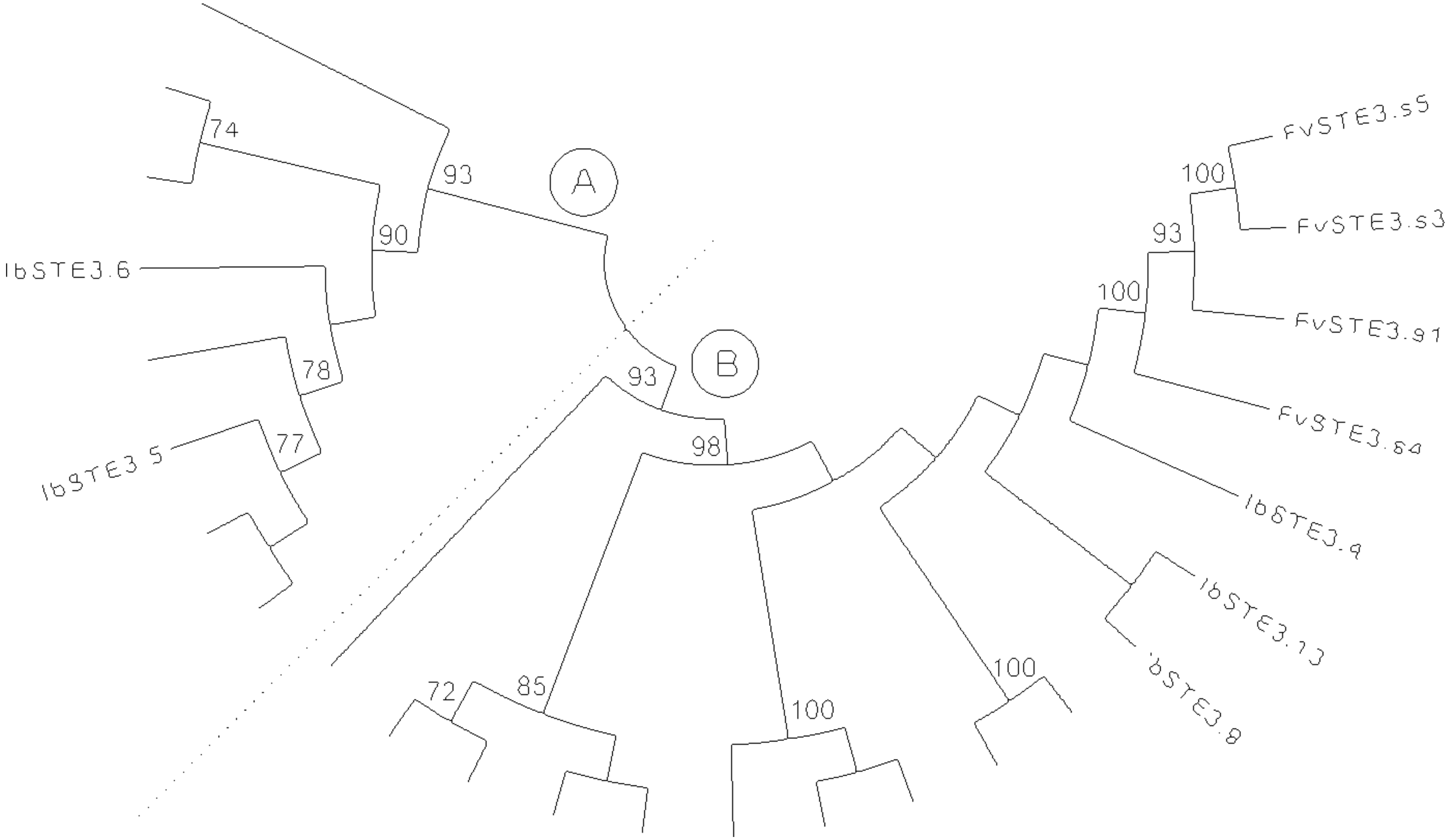
Our approach: automated!

- Faster than TreeSnatcher Plus
- Less picky about tree style than TreeRipper



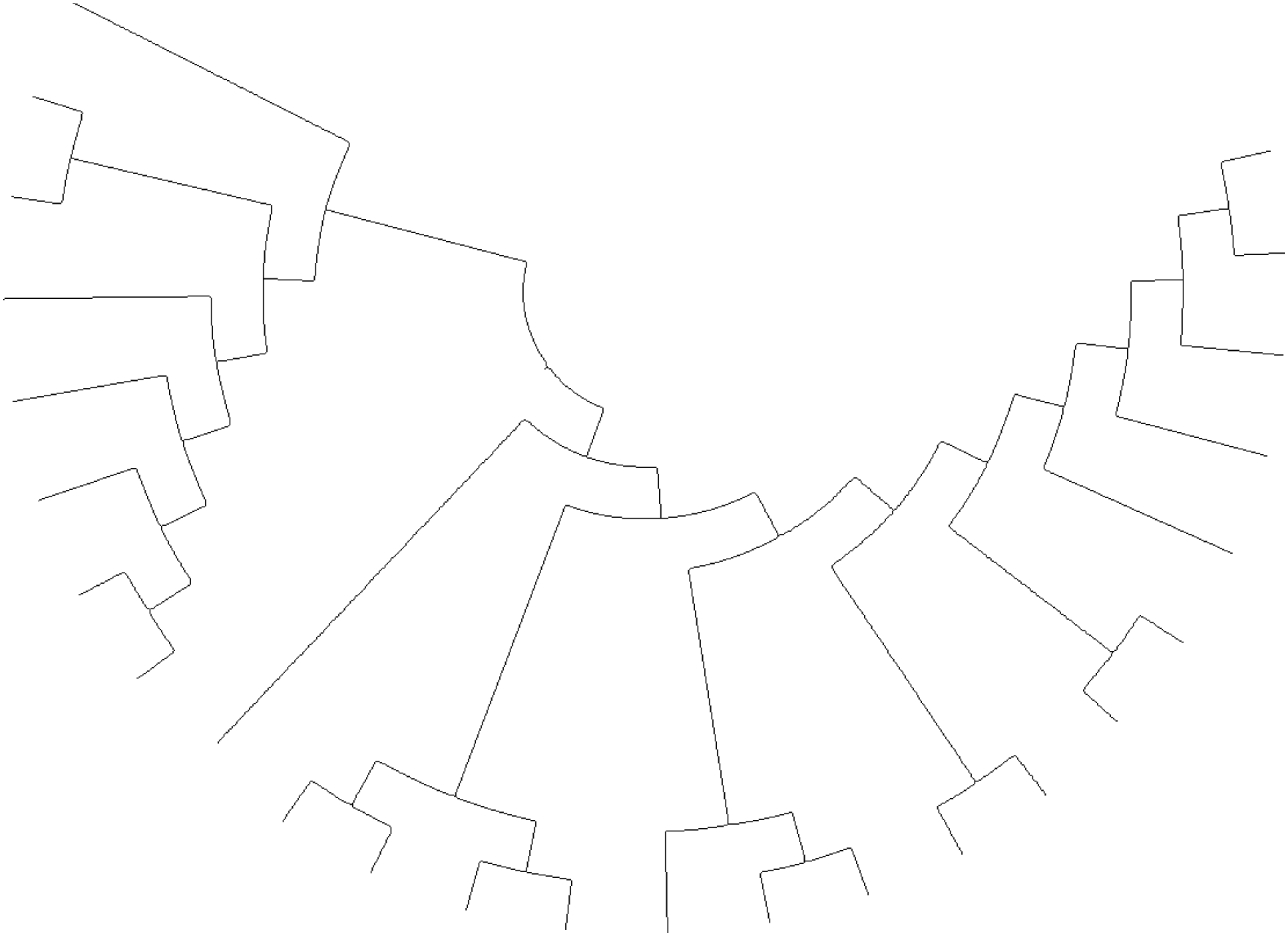
Stages 1&2 :

binarization (Black or White) & **thinning** (1 pixel width structures)



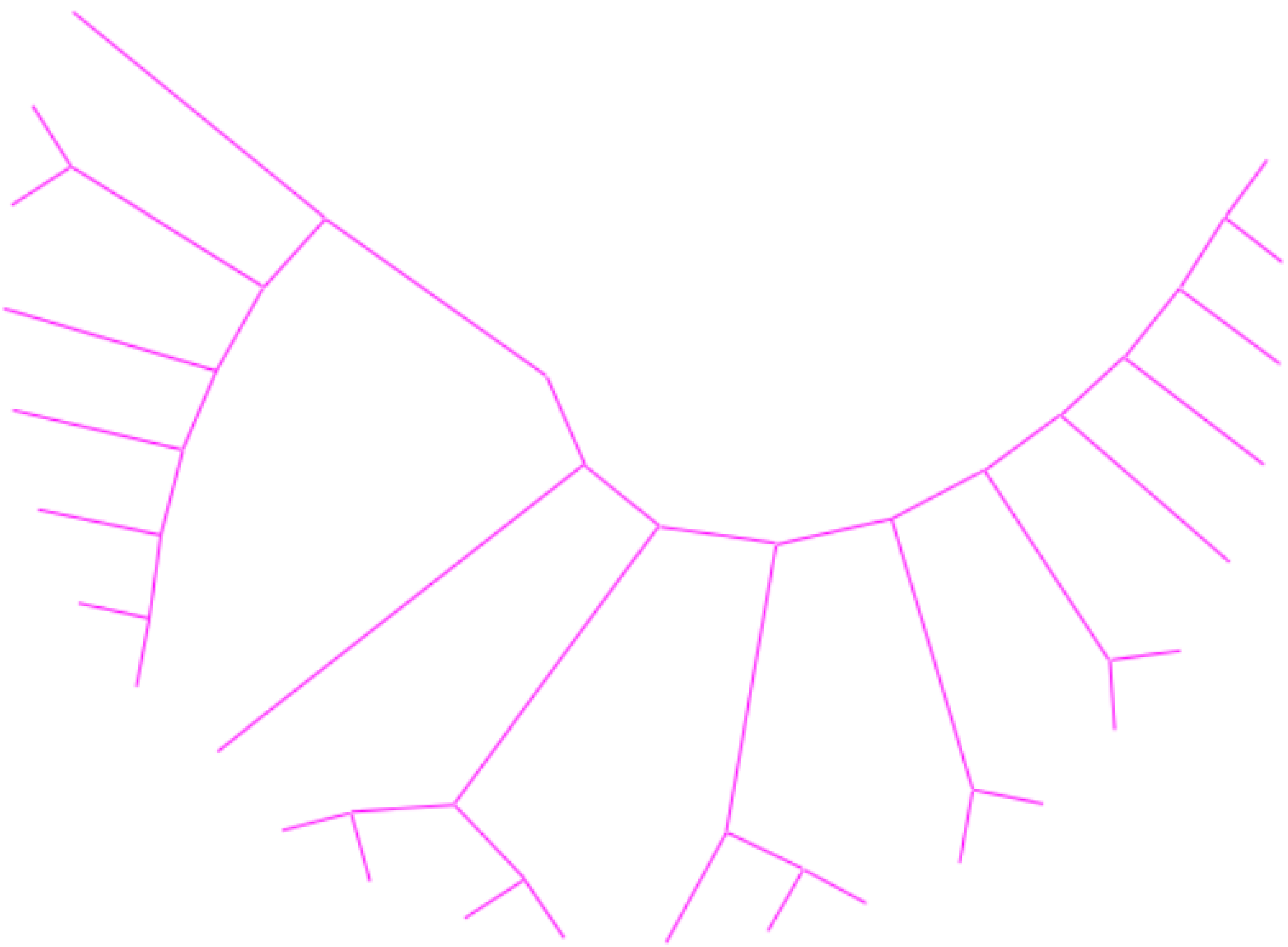
Stage 3 :

Assume largest 'pixel island' is the tree structure



Several stages later...

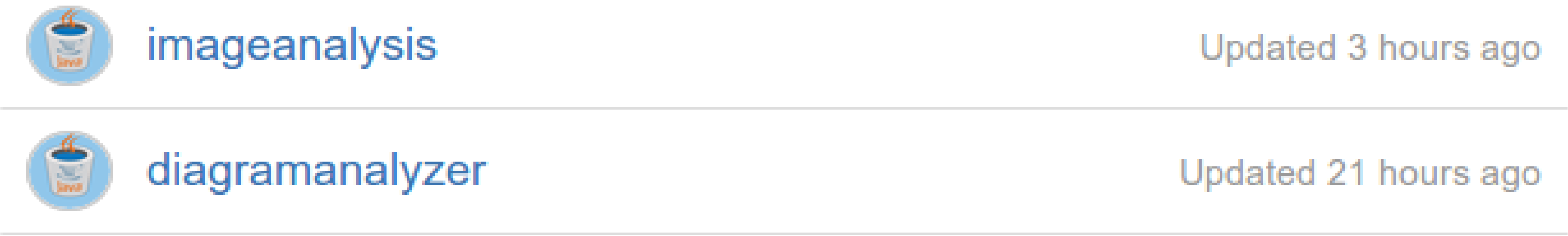
Re-draw / Re-use extracted data!



Still in very active development...

https://bitbucket.org/petermr/imageanalysis

https://bitbucket.org/petermr/diagramanalyzer

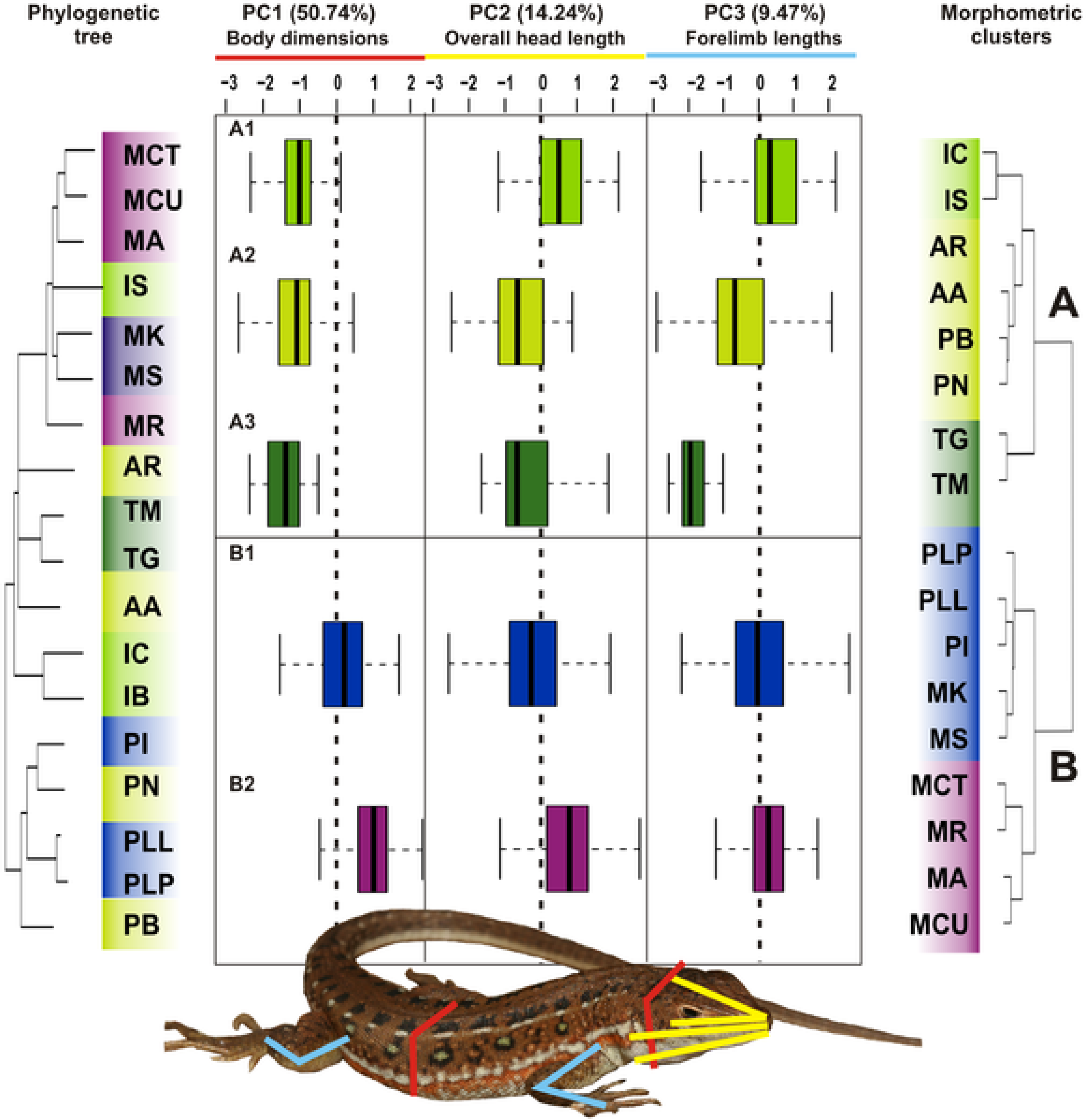


imageanalysis Updated 3 hours ago

diagramanalyzer Updated 21 hours ago

Java, Maven, Apache PDFbox, BoofCV,
Test-driven development, openly-licensed

Please stop publishing needlessly composite figures in *online-only* journals!!!


Phylogenetic tree
PC1 (50.74%) Body dimensions
PC2 (14.24%) Overall head length
PC3 (9.47%) Forelimb lengths
Morphometric clusters
A1
A2
A3
B1
B2
MCT
MCU
MA
IS
MK
MS
MR
AR
TM
TG
AA
IC
IB
PI
PN
PLL
PLP
PB
IC
IS
AR
AA
PB
PN
TG
TM
PLP
PLL
PI
MK
MS
MCT
MR
MA
MCU
A
B