**Phyloinformatic Literature Unlocking Tools**

Ross Mounce
Peter Murray-Rust
Matthew Wills

@rmounce
@petermurrayrust
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
Distribution of phylogen* articles 2000-2011

#1 is IJSEM
International Journal of Systematic & Evolutionary Microbiology

#5 PLOS ONE
(probably #3 now)

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 1\textsuperscript{st} 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

Make it a lot easier!

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

Chloroxylon

basic chromosome number X=9 or 10

ovary syncarpous; fruit indehiscent; endosperm 0

Aurantieae

chromones generally absent

Secretory cavities; seeds with tracheidal tegmen; mainly 1-3-seriate rays in the secondary xylem

Cneoroideae

chromones of ptaeroxylin group; secretory cavities absent in some groups

Rutoideae

Clade RTF

Some rights reserved
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 (IPS60464) after removing the covering crust and the infilling matrix. Model in (A) dorsal, (B) right lateral, (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

Java, Maven, Apache PDFbox, BoofCV, Test-driven development, openly-licensed
Please stop publishing needlessly composite figures in *online-only* journals!!!