# Tips for taxonomic cleaning with the TNRS

Brad Boyle

University of Arizona

9 January 2016

# Taxonomic cleaning

- Why bother?
- Taxonomic scrubbing applications
- General glitches and gotchas
- TNRS glitches and gotchas
- Pre-processing
- Post-processing
- Understanding the output

# Taxonomic cleaning: why bother?



**Hieronyma oblonga**

Widespread tropical tree

# Taxonomic cleaning: why bother?



**Hieronyma oblonga**

Widespread tropical tree

Synonym of *Hieronyma oblonga*, once thought to be endemic to Costa Rica

**Hieronyma poasana**

# Taxonomic cleaning: why bother?
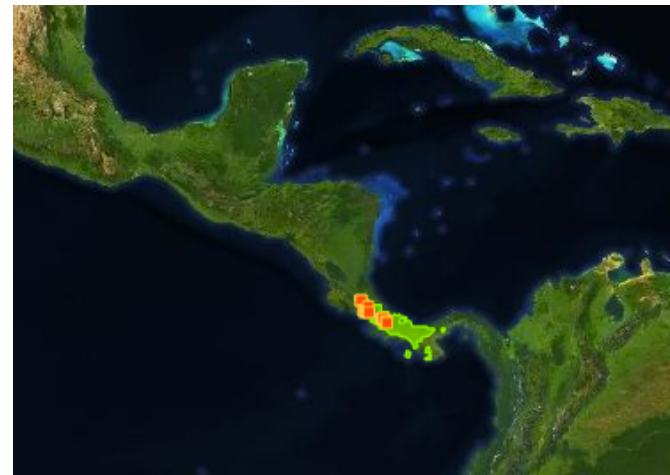
Widespread tropical tree

**Hieronyma oblonga**

Common misspellings of *Hieronyma oblonga*

**Hyeronima oblonga**
**Hieronima oblonga**

Synonym of *Hieronyma oblonga*, once thought to be endemic to Costa Rica

**Hieronyma poasana**

# Why bother?

**Table 5 Total names within two plant taxonomic databases before and after name resolution using the TNRS**

| Name source | Original names | After matching by TNRS | After matching & synonym conversion by TNRS |
|---|---|---|---|
| NCBI | 99743 | 97734 | 90142 |
| ITIS | 46483 | 45960 | 45025 |
| NCBI+ITIS (shared names) | 4412 | 19935 | 20670 |
| NCBI+ITIS (total unique names) | 141814 | 123759 | 114497 |

10% "bad" names

SOFTWARE                                                    Open Access

The taxonomic name resolution service: an online tool for automated standardization of plant names

Brad Boyle[1,2]*, Nicole Hopkins[2,3], Zhenyuan Lu[2,4], Juan Antonio Raygoza Garay[2,3], Dmitry Mozzherin[5], Tony Rees[6], Naim Matasci[1,2,3], Martha L Narro[2,3], William H Piel[7], Sheldon J Mckay[2,3,4], Sonya Lowry[2,3], Chris Freeland[8], Robert K Peet[9] and Brian J Enquist[1,10]

# Why bother?

**Table 5 Total names within two plant taxonomic databases before and after name resolution using the TNRS**

| Name source | Original names | After matching by TNRS | After matching & synonym conversion by TNRS |
|---|---|---|---|
| NCBI | 99743 | 97734 | 90142 |
| ITIS | 46483 | 45960 | 45025 |
| NCBI+ITIS (shared names) | 4412 | 19935 | 20670 |
| NCBI+ITIS (total unique names) | 141814 | 123759 | 114497 |

Overlap between databases only 3%!

**SOFTWARE**  **Open Access**

The taxonomic name resolution service: an online tool for automated standardization of plant names

Brad Boyle[1,2*], Nicole Hopkins[2,3], Zhenyuan Lu[2,4], Juan Antonio Raygoza Garay[2,3], Dmitry Mozzherin[5], Tony Rees[6], Naim Matasci[1,2,3], Martha L Narro[2,3], William H Piel[7], Sheldon J Mckay[2,3,4], Sonya Lowry[2,3], Chris Freeland[8], Robert K Peet[9] and Brian J Enquist[1,10]

# Why bother?

**Table 5 Total names within two plant taxonomic databases before and after name resolution using the TNRS**

| Name source | Original names | After matching by TNRS | After matching & synonym conversion by TNRS |
|---|---|---|---|
| NCBI | 99743 | 97734 | 90142 |
| ITIS | 46483 | 45960 | 45025 |
| NCBI+ITIS (shared names) | 4412 | 19935 | 20670 |
| NCBI+ITIS (total unique names) | 141814 | 123759 | 114497 |

400% increase in overlap

The taxonomic name resolution service: an online tool for automated standardization of plant names

Brad Boyle[1,2*], Nicole Hopkins[2,3], Zhenyuan Lu[2,4], Juan Antonio Raygoza Garay[2,3], Dmitry Mozzherin[5], Tony Rees[6], Naim Matasci[1,2,3], Martha L Narro[2,3], William H Piel[7], Sheldon J Mckay[2,3,4], Sonya Lowry[2,3], Chris Freeland[8], Robert K Peet[9] and Brian J Enquist[1,10]

# Taxonomic cleaning applications

- TNRS
  - (http://tnrs.iplantcollaborative.org/index.html)
- TaxonStand
  - http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2012.00232.x/full
- Global Name Resolver
  - http://resolver.globalnames.org/
- PlantMiner
  - http://www.plantminer.com/
- Many others...

# General architecture

- **Name parser**
  - Breaks up and classifies name components

Hieronima poasana Standley

Genus

Specific epithet

Authority
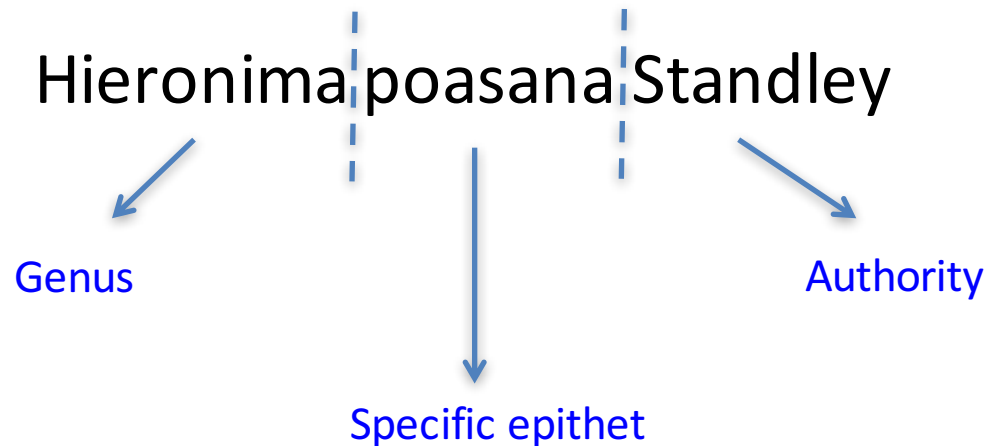
# General architecture

- **Name resolver**
  - Matches the name to reference database
  - Tries fuzzy matching if exact match fails

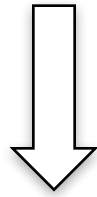Misspelled

Hieronima poasana Standley

Correct spelling
(as published)

Hieronyma poasana Standl.

# General architecture

- **Taxonomic status & synonym conversion**
  - Some applications do no do this last step

Synonym

Hieronyma poasana Standl.

Currently accepted name

Hieronyma oblonga (Tul.) Müll. Arg.

# Example workflow with TNRS API

- **Script**: tnrs_api_example.R
- **Steps:**

1. Extract the names
2. Turn into a string separated by commas
3. URL-encode and send to the TNRS API
4. Convert the returned JSON to data frame
5. Update your names

# Pros and Cons of TNRS API

- **Advantages**
  - Fast, simple, fully automated
- **Disadvantages**
  - Can't adjust all settings available in web interface
  - Uses Tropicos as only source
  - Can't take advantage of web interface to inspect results, choose alternative matches and research names
  - Can't access download options available in web interface
  - Parse-only option not available

# Example basic workflow with TNRS web interface

- **Script**: tnrs_gui_example.R
- **Steps**:
1. Extract names to CSV file with two columns: Unique ID & names
2. Upload to TNRS using bulk "Upload and Submit List" tab, checking box "My file contains an identifier as first column"
3. Adjust name processing settings and submit
4. Inspect results online, selecting alternate matches if appropriate
5. Download results, using options: Best matches only, Detailed results, UTF-8 format
6. Import TNRS results as tab-delimitted file
7. Remaining processing as for API

# Pros and Cons of TNRS Web Interface

- **Disadvantages**
  - Not fully automated
- **Advantages**
  - Can adjust name resolution settings
  - More name resolution sources
  - Use web interface to inspect results, choose alternative matches and research names
  - Select and download alternative matches on the fly
  - More download options, including "All matches" (useful if you don't like how TNRS chooses best match and want to script it yourself)
  - Parse-only more (useful for comparing part of original name to matched name)

# TNRS Tips & Gotchas

- Tip: Pre-pend family to name to prevent matching similar names in different families
- Gotcha: If you want to use The Plant List, *always* select TPL + ILDIS + GCC together
- Tip: Research any name where Taxonomic Status <> Accepted or Synonym
- Gotcha: Even accepted names can be wrong!

# Taxonomic Status

Taxonomic Status refers to the **Matched Name**

- **Accepted:** Good to go!
- **Synonym**: Good to go, as long as accepted name supplied
- **No opinion**: Could be good or bad name. RESEARCH IT
- **Invalid**: Never validly published. DON'T USE
- **Illegitimate**: Violates nomenclatural rules. DON'T USE
- **Rejected name**: Rejected by nomenclatural committee. DON'T USE
- **Misapplied name**: Commonly misapplied to the the wrong species. May or may not be correct. RESEARCH IT

# Even accepted names can be wrong!

| Name submitted | Tropicos | The Plant List |
|---|---|---|
| Henriettea fascicularis | =Henriettella fascicularis | =Henriettella fascicularis |
| Henriettea ramiflora | Accepted | Accepted |
| Henriettea succosa | Accepted | Accepted |
| Henriettella fascicularis | Accepted | Accepted |
| Henriettella tuberculosa | Accepted | =Henriettea tuberculosa |

Actually, all belong in *Henriettea*
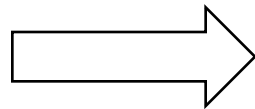
⟹

**Henrietteeae (Melastomataceae): A New Neotropical Berry-Fruited Tribe**

D. S. Penneys,[1,4] F. A. Michelangeli,[2] W. S. Judd,[3] and F. Almeda[1]

[1]Department of Botany, California Academy of Sciences, San Francisco, California 94118-4503, U. S. A.;
e-mail: dpenneys@calacademy.org; falmeda@calacademy.org.
[2]The New York Botanical Garden, 200th St. & Southern Blvd., Bronx, New York 10458, U. S. A.; e-mail: fabian@nybg.org.
[3]Department of Biology, University of Florida, Gainesville, Florida 32611, U. S. A.; e-mail: wjudd@botany.ufl.edu.
[4]Author for correspondence