# Taxonomic data quality in GBIF: a case study of freshwater insect groups

Joan E. Damerow, Patina K. Mendez, John Abbott, Petra Sierwald, Rudiger Bieler, Matthew J. Yoder, R. Edward DeWalt

Inaugural Digital Data in Biodiversity Research Conference,
Ann Arbor MI, June 6, 2017

The Field Museum

# Outline

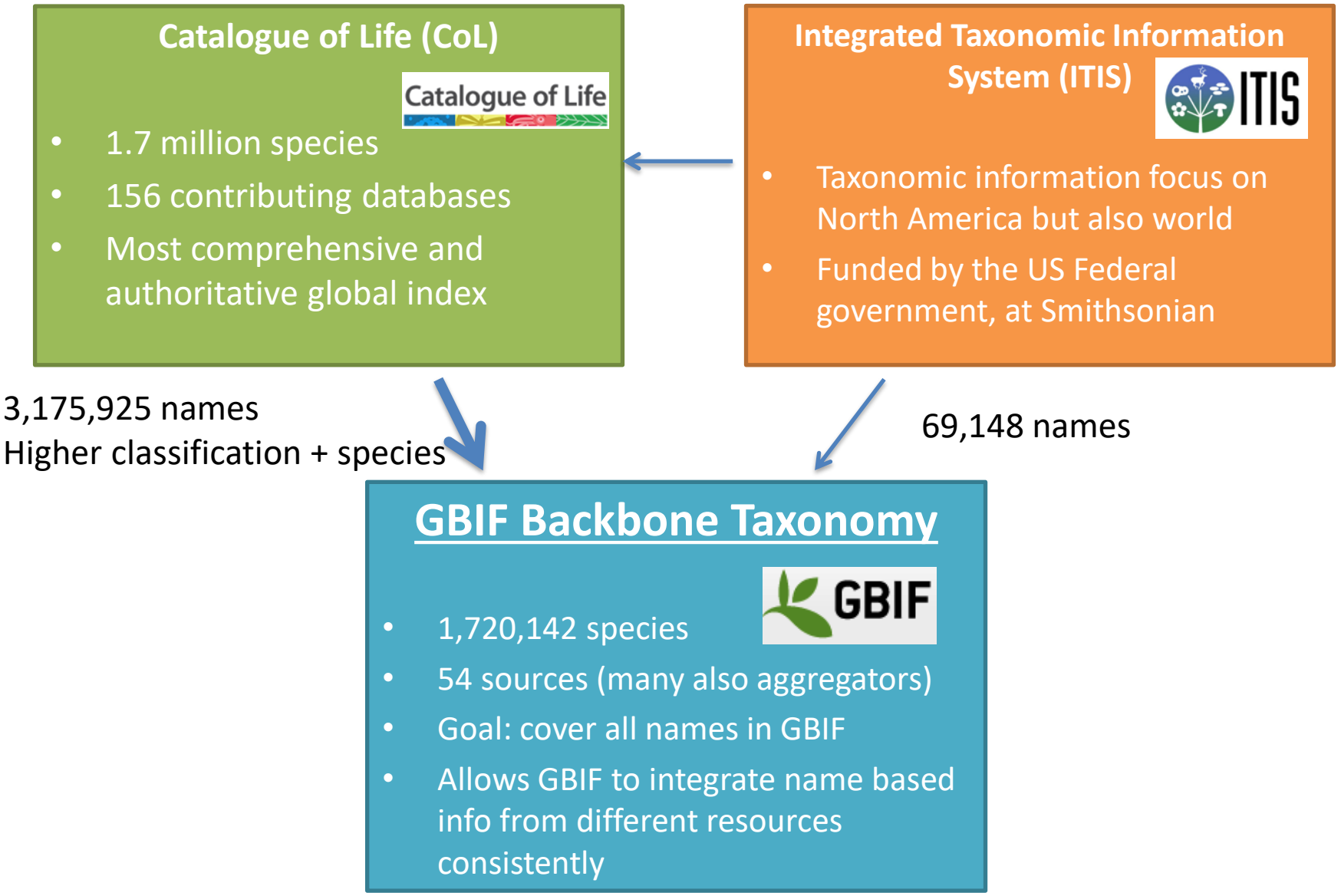
Focus on aquatic insect groups

**Questions:**

1. How do species list authors update taxonomic information in online aggregators?
2. How up to date is taxonomic info within the major aggregators?
3. How much does taxonomic information within aggregators affect GBIF occurrence records?

- Taxonomic aggregators
- GBIF Backbone Taxonomy
- Review of taxonomic sources to Catalogue of Life and ITIS
- Updating species lists and aggregators (survey)
- Prevalence of outdated names in GBIF records

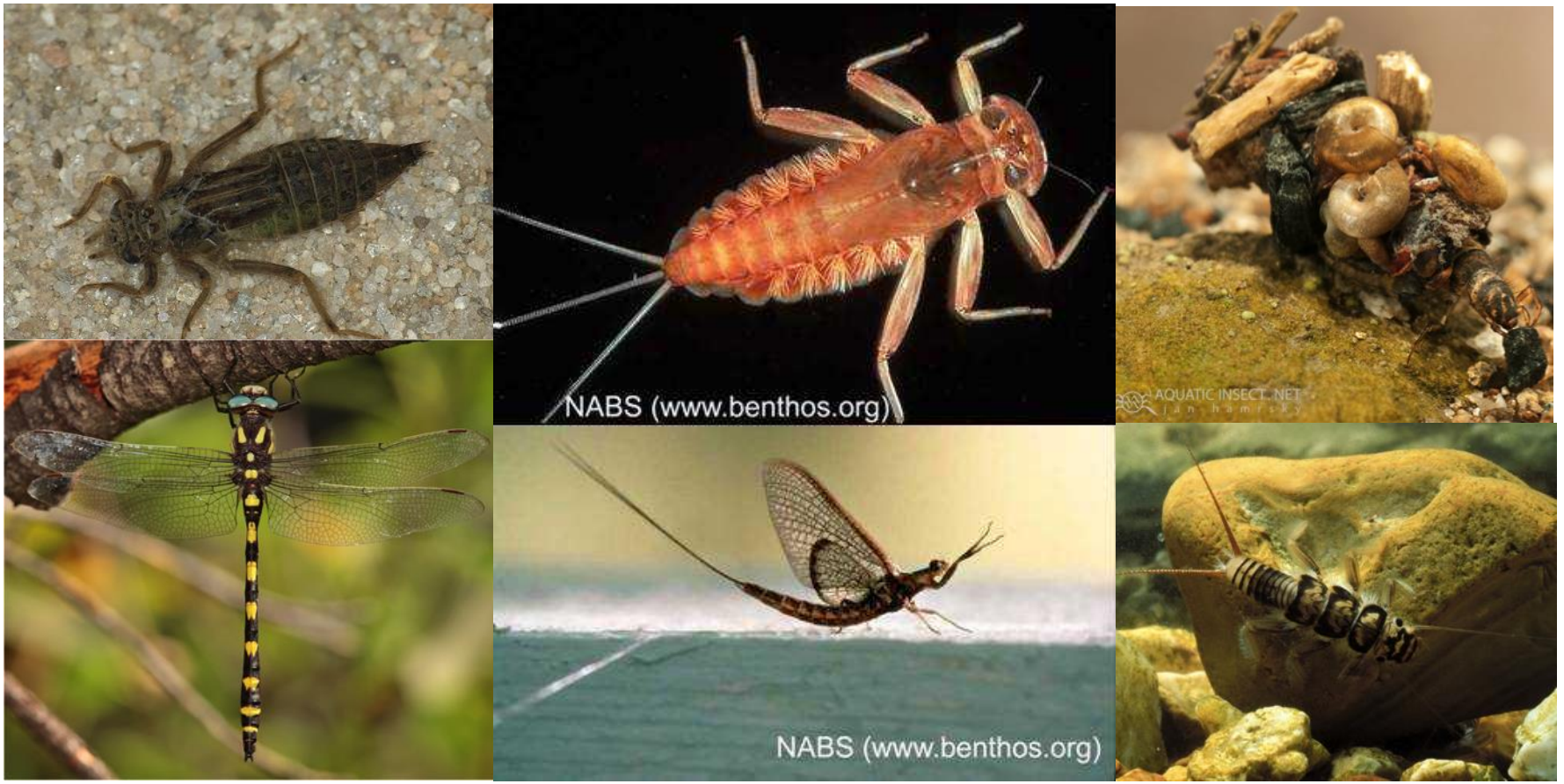
# Taxonomic data aggregators



# GBIF Backbone Taxonomy – Data Processing

- Harvest data from sources
- Interpret record: some data cleaning
- Issues: flag records with various problems or alterations during processing
- Taxonomy interpretation
- Match occurrence records to the backbone
  - Every occurrence assigned to taxonKey (matching taxon in backbone)
  - In case of homonyms or similar spelled names the service has a way to verify potential matches with higher taxonomy
  - If scientific name not yet part of GBIF backbone - can match the record to higher taxon (e.g. genus)
  - Taxon Match flags: fuzzy, high rank, no match
  - No flag for synonyms, but check ScientificName against Genus and Species (which will have accepted name)

# A case study of taxonomic information for dragonflies, mayflies, stoneflies and caddisflies



# Taxonomic sources - Odonata

**Odonata Central (Abbott)**

- 6261 valid species
- Online, searchable database
- Last Updated: March 2017

**World Odonata List (Schorr & Paulson)**

- 6241 valid species
- Online, non dynamic html list
- Last Updated May 2017

**New World Odonata (Garrison & von Ellenrieder)**

- Unknown valid species
- Word/PDF file
- Last Updated: March 2017

**Global Species Database of Odonata (Jan van Tol)**

- 5,912
- Website down/inaccessible
- Last Updated: Unknown

Updated 2004

Updated 2011

ITIS

Catalogue of Life

GBIF

# Taxonomic sources – Ephemeroptera

**Mayfly Central (McCafferty & Jacobus)**
- North and Central America
- Website, non-dynamic html
- Last Updated May 2017

**Ephemeroptera Checklist (Barber-James et al.)**
- Global
- Excel spreadsheet
- Last Updated 2015

**Updated 2009**

ITIS

**Updated 2013**

Catalogue of Life

FADA

GBIF

# Taxonomic sources

## Plecoptera

**Plecoptera Species File (DeWalt Et al.)**

- 3,769 valid species
- Website, searchable database SQL
- Last Updated: May 23, 2017

**Updated June 2017**

Catalogue of Life

GBIF

## Trichoptera

**World Trichoptera Checklist (John Morse)**

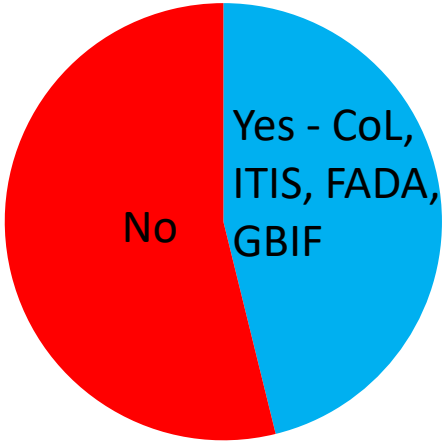
- 16,470 valid species
- Website, searchable database Filemaker
- Last Updated: July 2012, offline database updated daily

ITIS

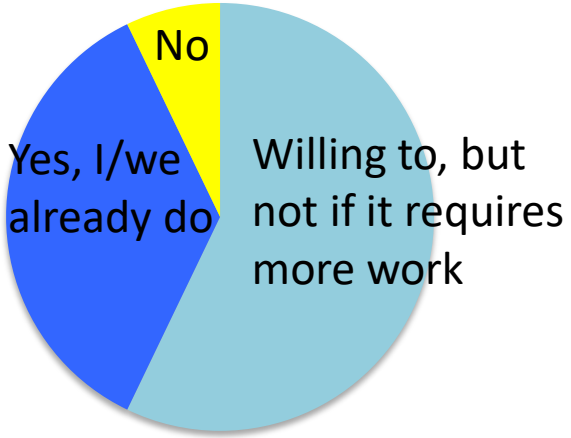
**Updated 2001**

# Pilot Survey –Updating Aggregators (n=14)

**Currently Update Aggregator(s)**




**Willingness to Update Aggregator(s)**




9 out of 14 respondents volunteer their time to update lists

Only 2 had financial or technical support

9/14 have updated within the last year

| | |
|---|---|
| Odonata | 3 |
| Plecoptera | 1 |
| Ephemeroptera | 2 |
| Trichoptera | 1 |
| Aquatic Coleoptera | 2 |
| Neuroptera/Megaloptera | 1 |
| Aquatic Diptera | 2 |
| millipedes Diplopoda | 1 |

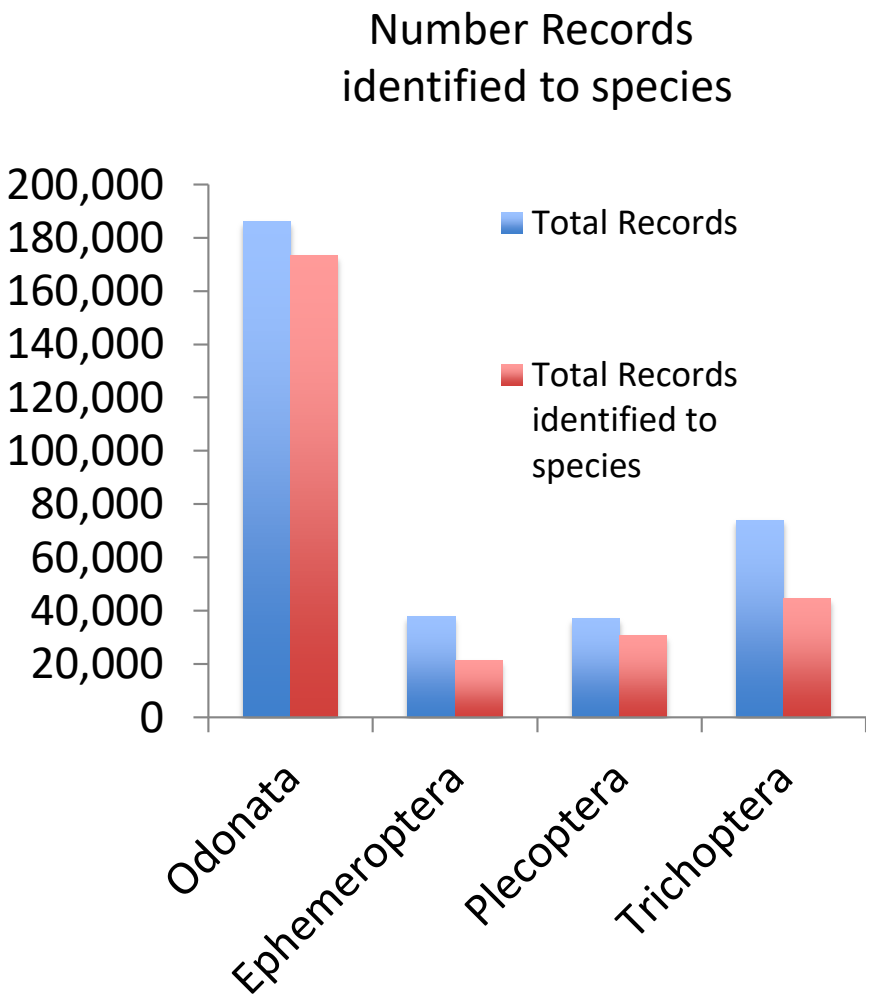
# General process for updating names in CoL and ITIS

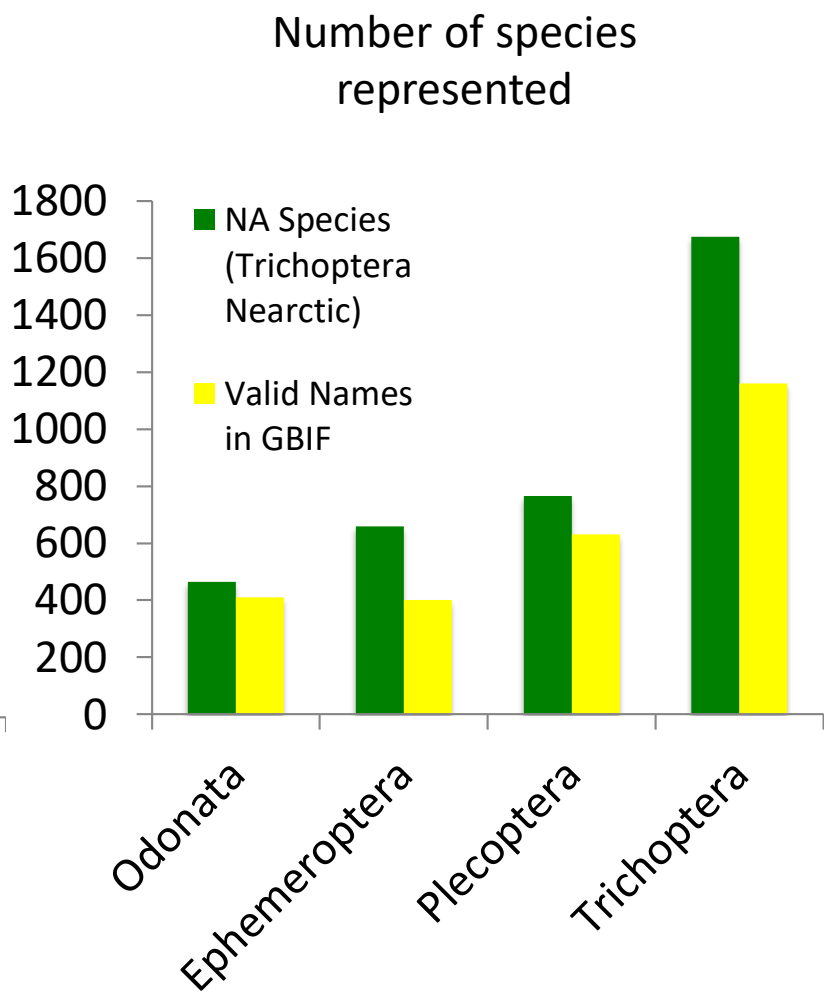
- Detailed instructions on website, and you can contact the editors
- Mandatory data fields: e.g. Accepted name, synonyms, references, classification above genus…
- New data: construct standardized file based on CoL/ITIS guidelines
- CoL is working on taking data in various formats (HTML, Word docs)
- Updates:
  - ITIS: Determine whether names are in ITIS using Compare Taxonomy tool
  - ITIS: Submit data for names not found in ITIS: data fields and standards, appropriate format
  - CoL: All changes submitted by contributing databases and passed to CoL through updates, when they remove old data and replace with new version
- Completed file sent to ITIS or CoL editor
- Peer Review
- New names assigned unique identifier

# Species File process for updating names in aggregators

- Semi-automated process: updates to CoL and GBIF twice a year (15 groups)
- Involves three staff at Species File Software Group and curators
- Communication with curators on what is needed for the next submission so they can resolve conflicts
- Integrity tests to look for data inconsistencies and omissions about a month before submission
- Summarize test results and call out specific data that needs attention
- Curators clean up that data over the next month, add recent papers that may not have been incorporated into their database yet, and review metadata
- For submittal- run integrity tests again and go through an export script to create format compatible for CoL and separately for GBIF
- Sent to CoL Executive Editor, and to GBIF

# GBIF Occurrence Records from USA – May 26, 2017



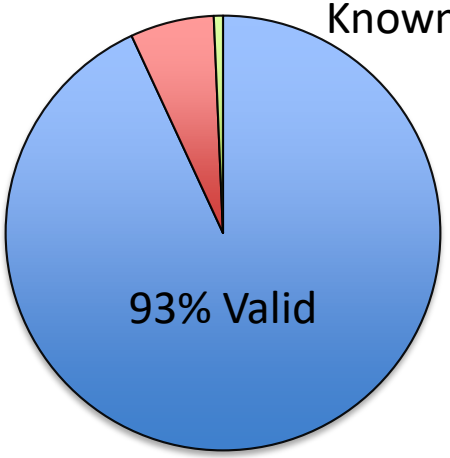



# Prevalence of Invalid Names – May 26, 2017

**GBIF Taxonomy Backbone**

**Last Updated 2011, different source**

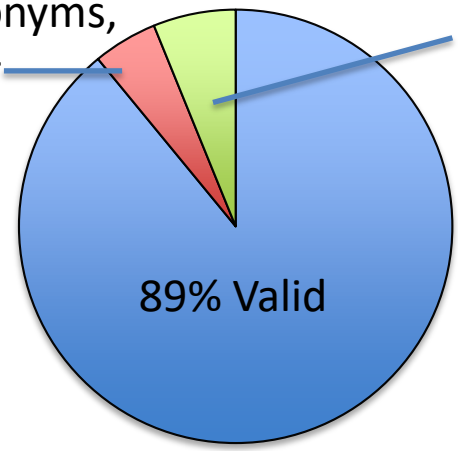
**Odonata**




**Ephemeroptera**

Known synonyms, other

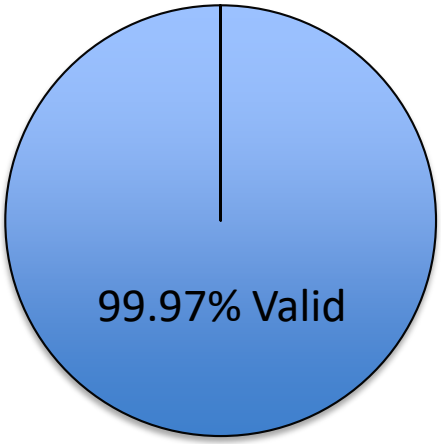
Invalid names-Unknown



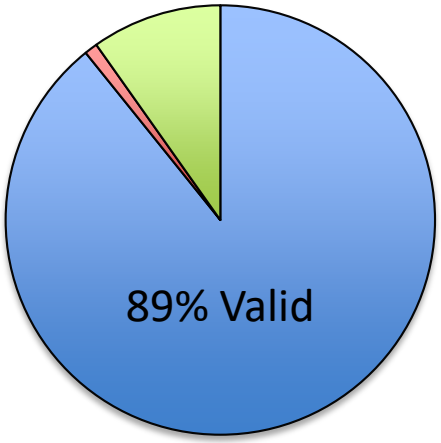

**Last Updated 2011**

**Last Updated May 2017**

**Plecoptera**




**Trichoptera**




**Last Updated 2001**

# Invalid names and GBIF Issue Flags

Taxonomic issues

- Taxon_Match_Fuzzy: Matching can only be done using a fuzzy, non exact match
- Taxon_Match_HigherRank: Matching can only be done on a higher rank and not the scientific name
- Taxon_Match_None: Matching cannot be done cause there was no match at all or several matches with too little information to tell them apart

| Taxa | # Invalid Records | # GBIF records w/ Taxa Issues | Invalid Records w/ Taxa Issues |
|---|---|---|---|
| Odonata | 11,980 | 4,367 | 70 |
| Ephemeroptera | 1,023 | 2,703 | 42 |
| Plecoptera | 10 | 1,317 | 0 |
| Trichoptera | 4,814 | 3,690 | 302 |

# Conclusions

- Available online data sources are rapidly changing
  - Better tools continue to be developed, but they are often difficult to find for outsiders
  - Not fully described in a single easy to find place
  - Data curators
- Prevalence of invalid names – we are already doing pretty well, depending on purpose
  - Data providers often (not always) update invalid names before submittal
  - GBIF data processing
  - GBIF sources are currently out of date, so you often need to check original sources
- Taxonomists need and can provide higher quality taxonomic information
- Taxonomic specialists are often not well supported and lack database skills
- More outreach and technical support is needed
- Goal: Up-to-date and comprehensive taxonomic information within online aggregators

# Thank You!

Joan Damerow
Bass Postdoctoral Fellow
Field Museum of Natural History
jdamerow@fieldmuseum.org

Photo by Ray Bruun

The Field Museum