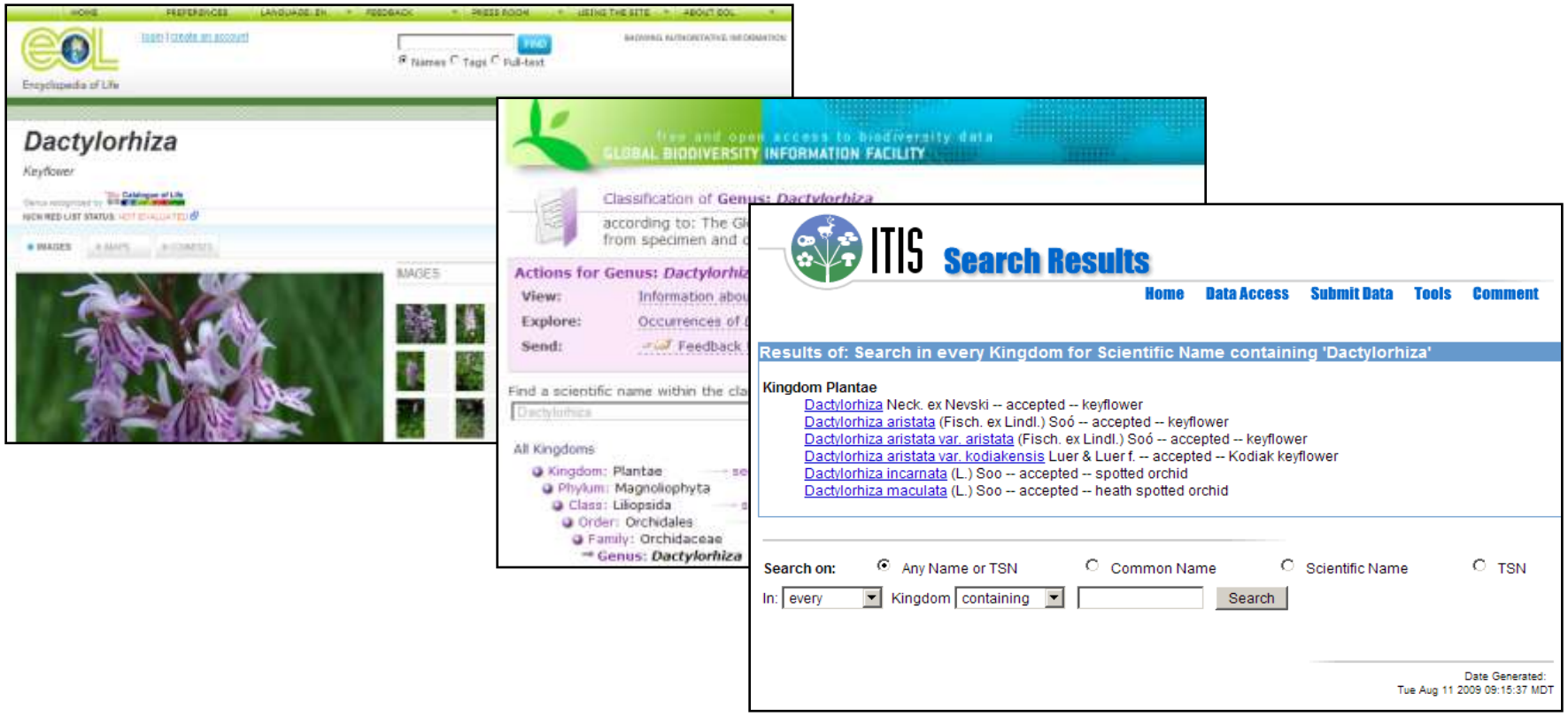
Towards building code compliance within biodiversity informatics

Alexey Shipunov,
Dmitry Mozzherin,
David Patterson
(Encyclopedia of Life)

Encyclopedia of Life



- Big Biodiversity Informatics “players” like EOL, ITIS and GBIF need to have access to all names of all organisms
- … and need set of rules to operate them

Encyclopedia of Life

Global Names Index

- This repository (Global Names Index, GNI, http://globalnames.org) is already exist as a part of Global Names Architecture collaborative project

Encyclopedia of Life

What taxonomists need

- Taxonomists need to from the full lists of names the ability to improve checklists and taxonomic descriptions at the levels of:
  - Common Latin misspellings;
  - The grammar specified by particular nomenclatural Code;
  - Analysis of homonyms;
  - Compliancy with Code rules

Encyclopedia of Life

Homonyms

- One of the biggest is homonyms problem. “True” homonyms are illegal, but still exist, especially among names of higher ranks.
- There are also hemiohomonyms belong to the scopes of different codes:
  - Our analysis revealed around 1100 “double-code” names, and even 22 “triple-code” names like *Rhodococcus*

Encyclopedia of Life

Homonyms

- In addition, there are “alien” names originated from different non-taxonomic sources but infiltrated to main databases:
  - Names of constellations like *Corona Borealis*;
  - Medical names like *Ossa metacarpalia;*
  - Syntaxonomical names, like *Glycerietum triflorae*

  and many others…

Encyclopedia of Life

Ambiregnal names

- Where bordes between codes are blurred, there is also a problem of names belong to several codes
  - The most widespread example are protistan names: *Euglena* could be regulated by both ICBN and ICZN

Encyclopedia of Life

Biocode

- The biggest problem, however, is the absence of unified code along with inconsistencies between existing codes
  - Five codes: ICBN, ICZN, ICNB, ICVCN, ICNCP plus emerging PhyloCode are incompatible in many aspects
  - … whereas the attempt to bring codes together (Biocode) was unsuccessful

Encyclopedia of Life

The idea

- Codes are rule sets, why not to translate codes into algorithms and then to software tools?
- Since every code is constantly changes, these tools would stay in versioning system which will accommodate all changes and proposals for all Codes, allowing to retrieve and use any version at any moment.

Encyclopedia of Life

How it could work

- Input:
  - User will choose one or multiply codes
  - Type or upload names, or supply text file containing names
- Output:
  - User will receive the output (Web page or text file) with all names commented for the compliancy with basic Latin grammar, selected code(s) (*Beroë* is not compatible with ICZN but compatible with ICBN), and homonymy (based on “warning lists”)

Encyclopedia of Life

Code and Code

23.4. The specific epithet, with or without the addition of a transcribed symbol, may not exactly repeat the generic name (such repetition would result in a tautonym).

Ex. 3. *"Linaria linaria"* and *"Nasturtium nasturtium-aquaticum"* are contrary to this rule and cannot be validly published.

Ex. 4. *Linum radiola* L. (1753) when transferred to *Radiola* Hill may not be named *"Radiola radiola"*, as was done by Karsten (1882), since that combination cannot be validly published (see Art. 32.1(c)). The next oldest name, *L. multiflorum* Lam. (1779), is illegitimate, being a superfluous name for *L. radiola*. Under *Radiola*, the species has been given the legitimate name *R. linoides* Roth (1788).

```
1. if ($generic_name == $species_ephithet)
2.     {
3.     $name["icbn_validly_published"] = FALSE;
4.     }
```

- This will not substitute codes but provide a common framework for all of them

Encyclopedia of Life

How it is working now: Name Parser


```
---
- scientificName:
    canonical: Drosera obovata
    positions:
      "26":
      - author_word
      - 36
      "0":
      - genus
      - 7
      "18":
      - author_word
      - 23
      "10":
      - species
      - 17
    verbatim: Drosera x obovata Mert. & W.D.J.Koch
    details:
    - genus:
        epitheton: Drosera
      species:
        epitheton: obovata
        basionymAuthorTeam:
          authorTeam: Mert. & W.D.J.Koch
          author:
          - Mert.
          - W.D.J.Koch
        authorship: Mert. & W.D.J.Koch
    normalized: "Drosera \xC3\x97 obovata Mert. et W.D.J.Koch"
    hybrid: true
    parsed: true
```


Encyclopedia of Life

How it is working now: NameLink (1)

HOME | PREFERENCES | LANGUAGE: EN | FEEDBACK | PRESS ROOM | USING THE SITE | ABOUT EOL

Encyclopedia of Life

login | create an account

FIND

Names Tags Full-text

SHOWING AUTHORITATIVE INFORMATION

NameLink

NameLink is a service offered by EOL that allows you to submit a webpage address and have the taxon names within the page automatically identified and linked up to projects which have information about those names. Developers can easily embed this functionality within their own webpages by using the NameTag API, as described in the documentation. Note that this service may produce some unexpected results and may not yet work correctly in all browsers. Some of the unexpected results are due to the difficulty in accurately recognizing and disambiguating taxon names.

To see NameLink in action, you can use the form below to submit any webpage you wish. Note that it may take a few seconds for the names to begin to be recognized after you submit the webpage.

Select project that should be linked to: Global Names Index

☑ Add common names if possible
☑ Show project logos next to links
URL: http://herba.msu.ru/shipunov/knapweed/stoebe_ochsmann.htm

Click one of the following example URLs to paste it into the box above:
A Bioline article and abstract: http://www.bioline.org.br/abstract?id=fb95003
A list of endangered species: http://www.fws.gov/Endangered/1966listing.html

NameLink

Encyclopedia of Life

How it is working now: NameLink (2)

1. Ray florets missing, flowers cream-coloured, rarely purple, capitula (widest lower part) ca 2-6 mm wide, phyllary appendages always with distinct terminal spine, achenes small, 2-3 mm, pappus 0 (at most rudimentary) ... C. diffusa

– Ray florets present, capitula ca 3.5-11 mm wide ... 2

2. Flowers whitish or purple (intermediate colours occur), capitula ca 3.5-8 mm wide, phyllary appendages light to dark, very variable, terminal spine 0 to well developed, length of achenes variable, 2.5-3.5 mm, pappus missing or up to 0.6 mm ... C. x psammogena (C. diffusa x C. stoebe)

– Flowers purple, capitula over 5 mm wide, terminal spine 0 ... 3

3. Pappus ca 0.5-1 mm (sometimes 0?), phyllary appendages big, black, with 8-11 cilia per side (endemic in SE Europe) ... C. stoebe subsp. serbica (= C. affinis, C. tartarea)

– Pappus ca 1-2.5 mm, appendages dark brown to black, with 4-10 cilia per side ... 4

4. Plants normally single-stemmed, monocarpic, capitula ca 6.5-11 mm wide, phyllary appendages with 6-10 cilia per side, diploid (2n = 18), Europe, absent in North America ... C. stoebe subsp. stoebe (= C. rhenana, C. paniculata, C. maculosa)

– Plants normally many-stemmed, perennial, polycarpic, capitula ca 5-8 mm wide, phyllary appendages [illegible] cilia per side, bracts often tinged with dark violet, tetraploid (2n = 36), invasive in Europe and North America ... C. stoebe subsp. micranthos (= C. micrantha, C. biebersteinii)

1. Ray florets missing, flowers cream-coloured, rarely purple, capitula (widest lower part) ca 2-6 mm [illegible] appendages always with distinct terminal spine, achenes small, 2-3 mm, pappus 0 (at most rudimentary) ... C. diffusa

– Ray florets present, capitula ca 3.5-11 mm wide ... 2

2. Flowers whitish or purple (intermediate colours occur), capitula ca 3.5-8 mm wide, phyllary appendages light to dark, very variable, terminal spine 0 to well developed, length of achenes variable, 2.5-3.5 mm, pappus missing or up to 0.6 mm ... C. x psammogena (C. diffusa x C. stoebe)

– Flowers purple, capitula over 5 mm wide, terminal spine 0 ... 3

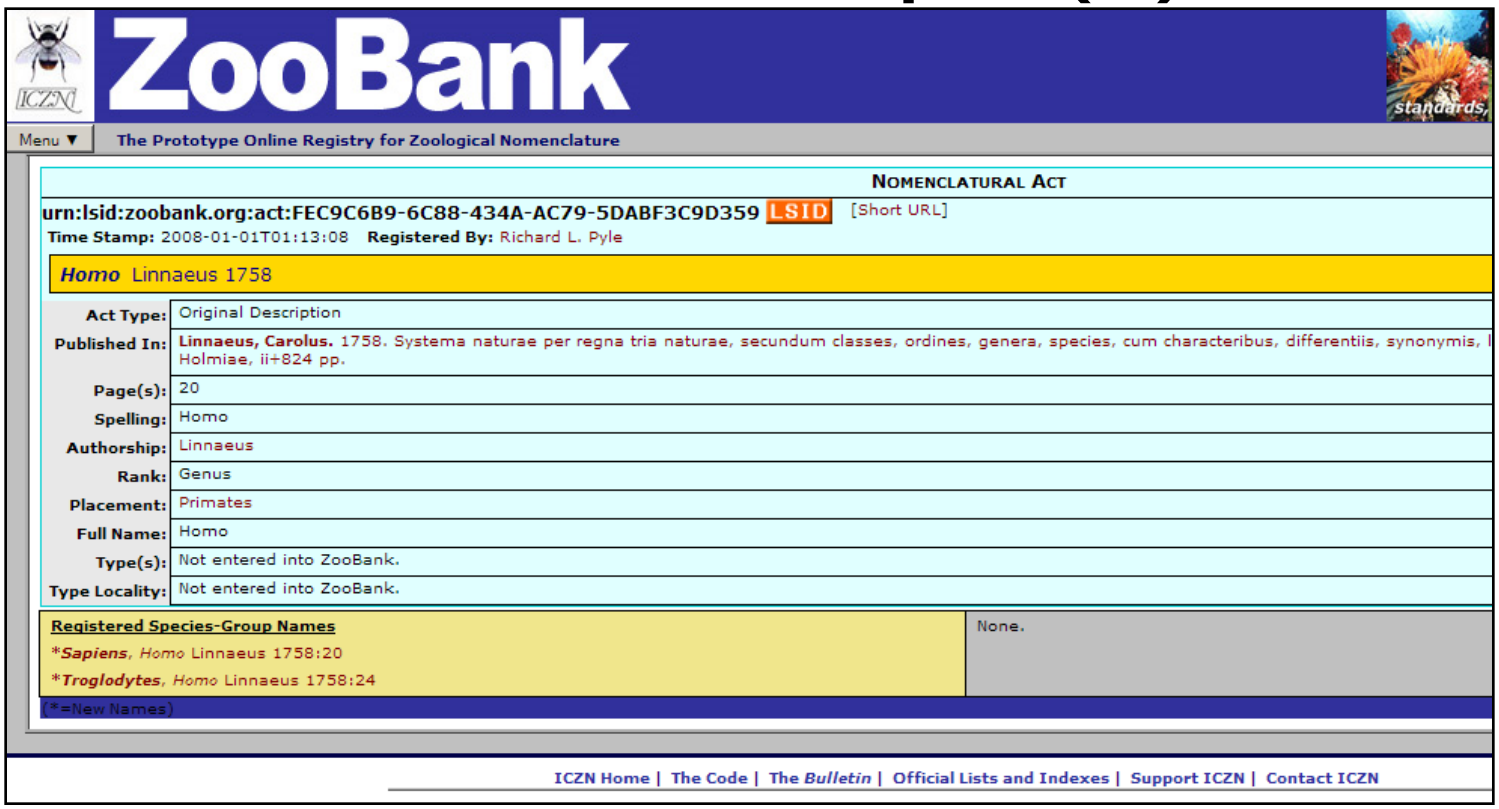
3. Pappus ca 0.5-1 mm (sometimes 0?), phyllary appendages big, black, with 8-11 cilia per side (endemic in SE Europe) ... C. stoebe subsp. serbica (= C. affinis, C. tartarea)

– Pappus ca 1-2.5 mm, appendages dark brown to black, with 4-10 cilia per side ... 4

4. Plants normally single-stemmed, monocarpic, capitula ca 6.5-11 mm wide, phyllary appendages with 6-10 cilia per side, diploid (2n = 18), Europe, absent in North America ... C. stoebe subsp. stoebe (= C. rhenana, C. paniculata, C. maculosa)

– Plants normally many-stemmed, perennial, polycarpic, capitula ca 5-8 mm wide, phyllary appendages with ca 4-7 cilia per side, bracts often tinged with dark violet, tetraploid (2n = 36), invasive in Europe and North America ... C. stoebe subsp. micranthos (= C. micrantha, C. biebersteinii)

Future steps (1)



ZooBank

Menu ▼ The Prototype Online Registry for Zoological Nomenclature

NOMENCLATURAL ACT

urn:lsid:zoobank.org:act:FEC9C6B9-6C88-434A-AC79-5DABF3C9D359 LSID [Short URL]

Time Stamp: 2008-01-01T01:13:08 Registered By: Richard L. Pyle

Homo Linnaeus 1758

Act Type:	Original Description
Published In:	**Linnaeus, Carolus.** 1758. Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, l Holmiae, ii+824 pp.
Page(s):	20
Spelling:	Homo
Authorship:	Linnaeus
Rank:	Genus
Placement:	Primates
Full Name:	Homo
Type(s):	Not entered into ZooBank.
Type Locality:	Not entered into ZooBank.

Registered Species-Group Names	None.
***Sapiens**, Homo* Linnaeus 1758:20	
***Troglodytes**, Homo* Linnaeus 1758:24	

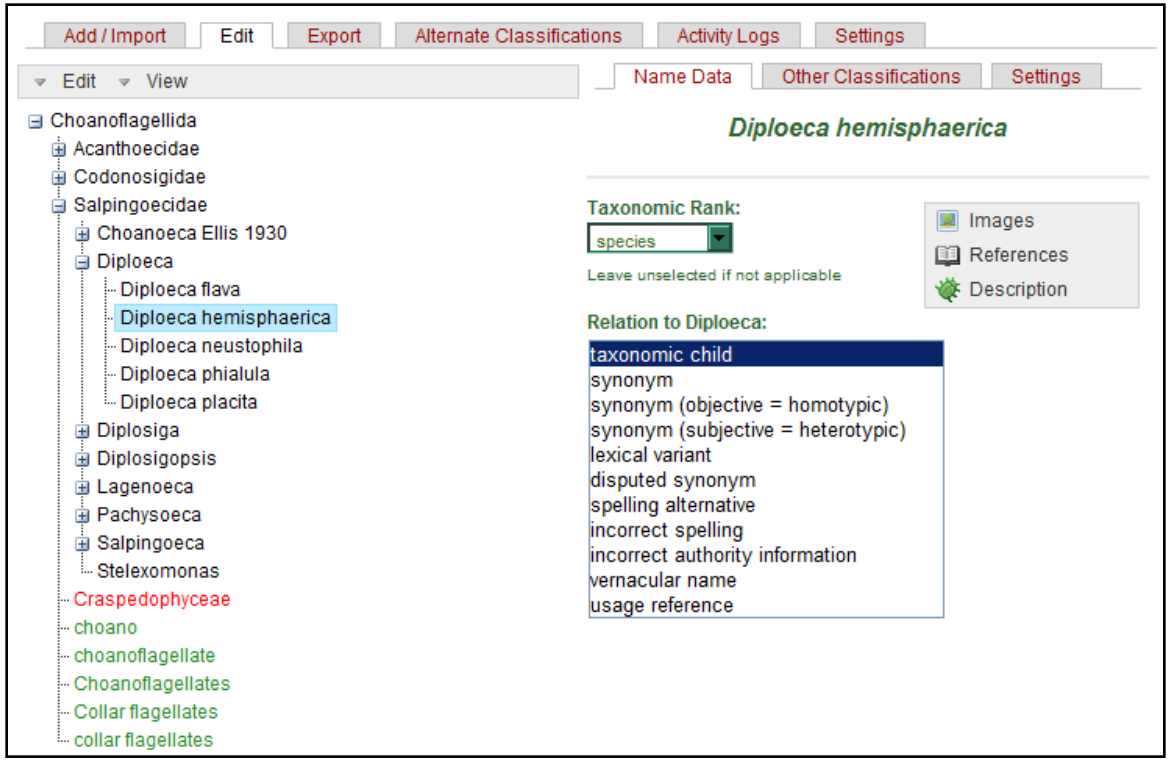
(*=New Names)

ICZN Home | The Code | The *Bulletin* | Official Lists and Indexes | Support ICZN | Contact ICZN

- With the emerging name registration for ICZN in ZooBank, these tools could become an essential part of registration process

Encyclopedia of Life

Future steps (2)



- The emerging taxonomic editor (NSF proposal, in review) will have these tools as a module

Encyclopedia of Life

Future steps (3)

- Software tools explaining here will create a base ground for further unification of names, where
  - **Resolving of hemiohomonyms**, like *Rhodococcus* (z), *Rhodococcus* (b) and *Rhodococcus* (a)

  and

  - **Unification of higher-ranked names**, like [5]*Cycas* for “Classis Cycadopsida”, or [6]*Sagitta* for “Phylum Chaetoghnatha”

  are extremely wanted

Encyclopedia of Life

Acknowledgements

- Patrick Leary (EOL)
- David Shorthouse (EOL)
- Thomas Orrell (ITIS)
- David Remsen (GBIF)