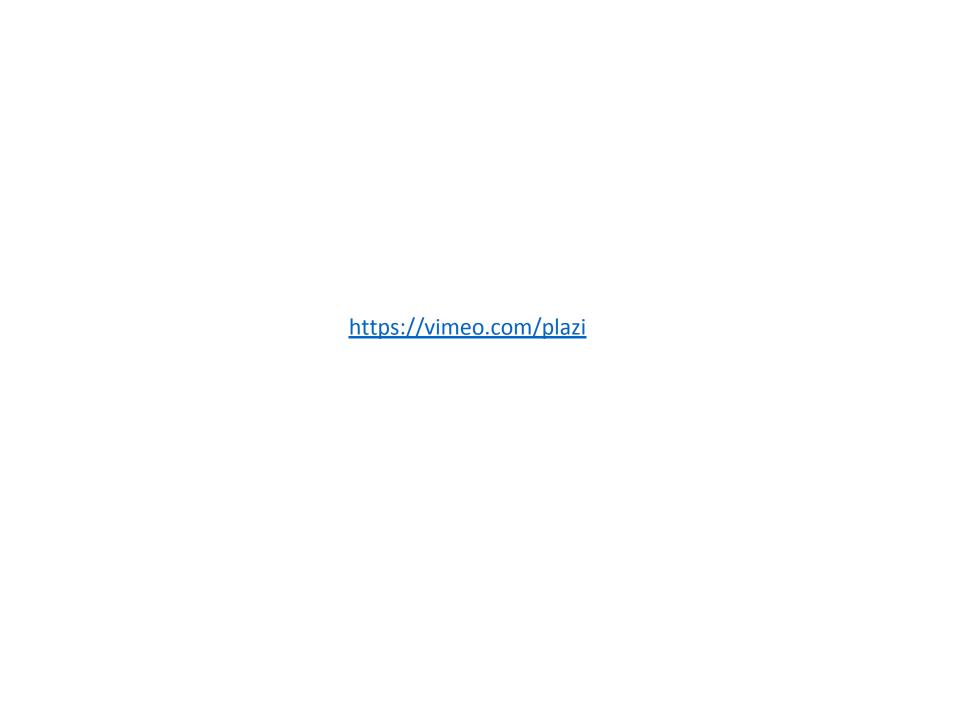# Expanding the workflow from scholarly published data to GBIF

Donat Agosti

GBIF Global Nodes Meeting, Leiden, October 19, 2019

https://vimeo.com/plazi

The scope

- Ca 500M pages published, including the scholarly taxonomic knowledge about the Earth' taxa
- Each year new data is published in over 1.000 journals and books, including ca 17,000 new animal species
- Data published with implicit, well defined relation to species (taxonomic treatments)
- Data in publications are micro-citation networks with external and internal links
- All scientific names are published, including their history (synonomies)
- Scientific taxonomic names are citations of data in published records
- Tens of millions of treatments are referenced in the literature and databases
- Billions of facts about species can be derived and linked to the source publication
- Data in taxonomic publications are in the public domain
- Text and data mining tools exist to extract, annotate and make accessible data from publications
- The Biodiversity Literature Repository is in place making data (Figures, Taxonomic Treatments) findable, accessible, interoperable, reusable)
- A fully automated workflow to liberate data exists from a journal article to the reuse of data in GBIF
- Publications are in place providing semantically enhanced data.

Scholarly published biodiversity data

Data in taxonomic publications is very rich

Data is not copyrighted and thus freely accessible and reusable.

Scholarly publication PDF Print

Data in a publication:

⭐ Metadata
  - Journal
  - Authors
  - Affiliations
  - ORCID

⭐ Figures, caption, figure citations
⭐ Tables, caption, table citations
- Bibliographic references, bibliographic reference citations

- Materials and methods
- Methodologies
- Collection codes
- Measuring standards

- Taxonomy
- Synonymic lists
⭐ Identification keys
- **Taxonomic treatments**

Data in **taxonomic treatments**
⭐ Taxon name
- Taxonomic name labels
- Nomenclatural acts
⭐ Thematic Sections
- Descriptions with traits
⭐ Typed citations of previous treatments (synonymy)
⭐ Distribution
- Behavior
⭐ **Materials citations**
- Relations with other taxa

Data in **materials citations**
- Taxon name
⭐ country
⭐ locations
⭐ Geo-coordinates
⭐ Collecting dates
⭐ Collector
⭐ Collecting Methods
⭐ Habitat
⭐ Microhabitat
⭐ Collected from
⭐ Identifying persons
⭐ Collection codes
⭐ Specimen codes
- Specimen counts
⭐ DNA Accession codes
⭐ Type status

⭐ Data liberated from publications and used by GBIF
Blue List of data in taxonomic publications: http://plazi.org/activities/blue-list/

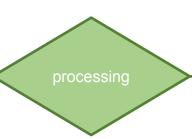# Plazi scholarly published data liberation and dissemination workflow

Publication
* PDF
* HTML
* XML

processing

Open FAIR Data
* Article metadata
* Figures
* Treatments

dissemi-nation

Re-use

article with enriched metadata

figure

treatment

Biodiversity Literature Repository

dataset

treatments

figures

occurrence

Unique names, removing duplicate treatments with the same name:
all: 218.785
genus: 40.750
species: 165.634
infraspecies: 6.044

Plazi provided names, not found anywhere else in the backbone sources:
all: 48.744
genus: 1.861
species: 43.770
infraspecies: 3.113

Plazi mediated treatments in GBIF
286K

Coming

* Multiplying the data input to GBIF
* collaboration with Publishers to streamline publishing of observation records
* adding synonyms
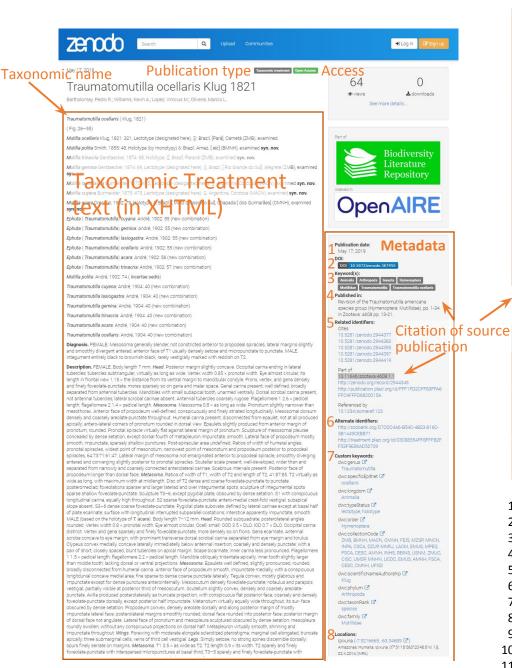

Collaboration

* Make it an imperative to have your treatments in GBIF
* Convince the user of your data to cite specimens with persistent digital identifiers in publications
* Use the treatment persistent identifiers in your work

Thank you!

Taxonomic treatments are data that can be made FAIR

Metadata

1. Original publication date of the article including the treatment
2. Digital object identifier
3. Keywords free text
4. Bibliographic citation of source publication
5. Related identifiers
6. Alternative Identifiers
7. Custom keywords using domain specific vocabularies
8. Location of material citations (location, geo-coordinates, verbatim text)
9. Dates of material collected
10. Communities
11. License

https://sandbox.zenodo.org/record/367955#.XYhu5igzbAQ

BLR: Biodiversity Literature Repository

Existing
- implemented
- planned

Suggested
DiSSCo/ICEDIG