PLAZI
TAKING CARE OF FREEDOM

Doi: 10.5281/zenodo.6574848

Discovering known biodiversity

May 24, 2022

Empowering Biodiversity Research II conference

Africa Museum, Tervuren

Donat Agosti

ORCID: 0000-0001-9286-1200

Plazi, Switzerland

SIB LS
SWISS INSTITUTE OF BIOINFORMATICS
LITERATURE SERVICES

swissuniversities

101007492

ARCADIA

BiCIKL

eBioDiv

Biodiversity Literature Repository

TreatmentBank

Quiz

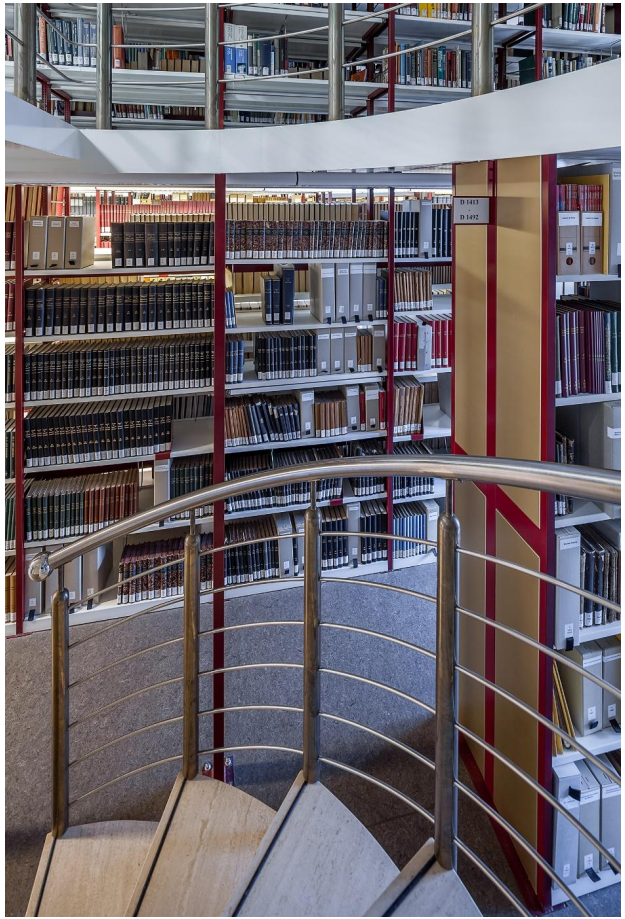
What is the taxonomic name of the Puer tea?

Provide as evidence:

- Taxonomic name
- Digital copy of the protologue of the taxon
- Evidence of holotype
- Digital copies of the protologues of five synonyms
- Digital copies of the documentation of the synonymy
- Digital copy of the phylogenetic position of the taxon
- Time sheet including processing time

Prize:

The challenge



Known biodiversity knowledge

- Empiric science: all results published
- 500,000,000+ printed pages
- >> 1,000 journals publishing taxonomic content
- > 1,900,000 species described
- > 20,000,000+ taxonomic treatments
- approx. 18,000 new species discovered / year
- >> Millions of specimens identified by specialists (material citations)
- Billions of facts

BUT: only ca. 10-30% of the knowledge is digital. Most is “unknown known knowledge”, not Digital Accessible Knowledge (DAK)

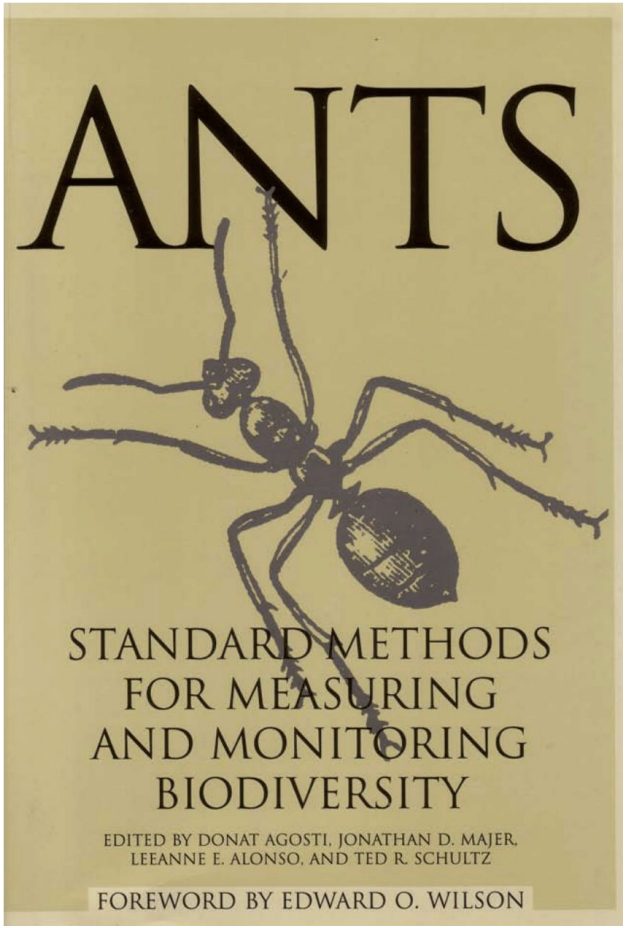
Take home points:

- Nothing in taxonomic publications makes sense except in the light of taxonomic treatments
- An identification of a specimen is only complete with a link to a taxonomic treatment
- A material citation in a publication is the gateway to the knowledge about the specimen
- Automate as much as possible, curate as much as needed
- Be aware of the power and use of your data beyond domain: make use of semantic publishing

Rio Earth Summit 1992: Biodiversity crisis, 30 years ago

How many species do we lose?
How many species do we know?
How many species are on Earth?
What do we know about the species?

Measuring and monitoring biodiversity in the 90ties



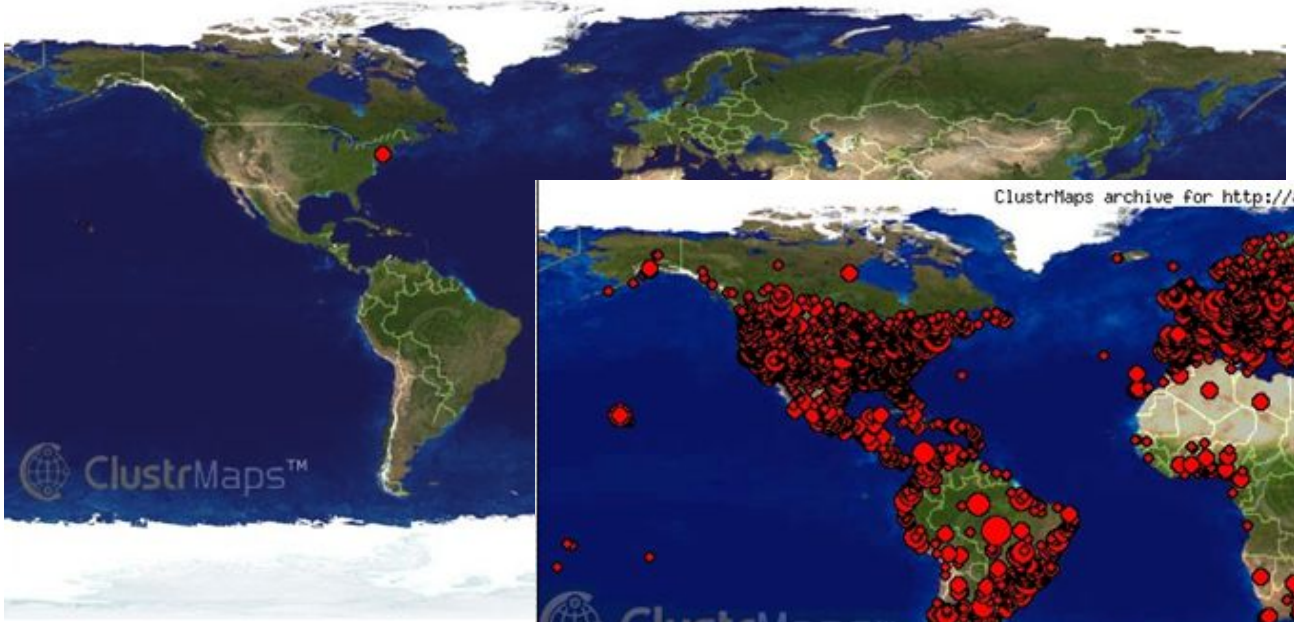
Basic questions in any monitoring program:

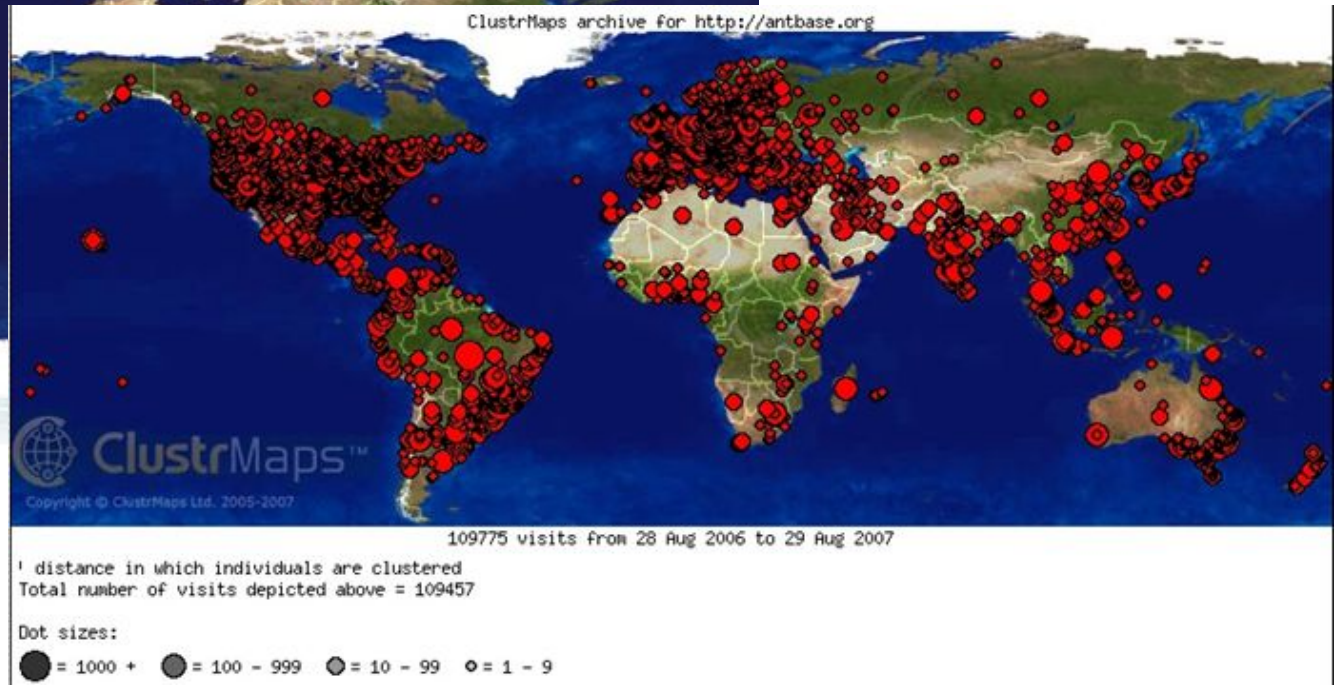
- What taxon is it?
- How does it fit into the phylogeny and classification?
- What do I know about it?
- What is its distribution?

https://doi.org/10.5281/zenodo.11736

The impact of WWW in 90ties

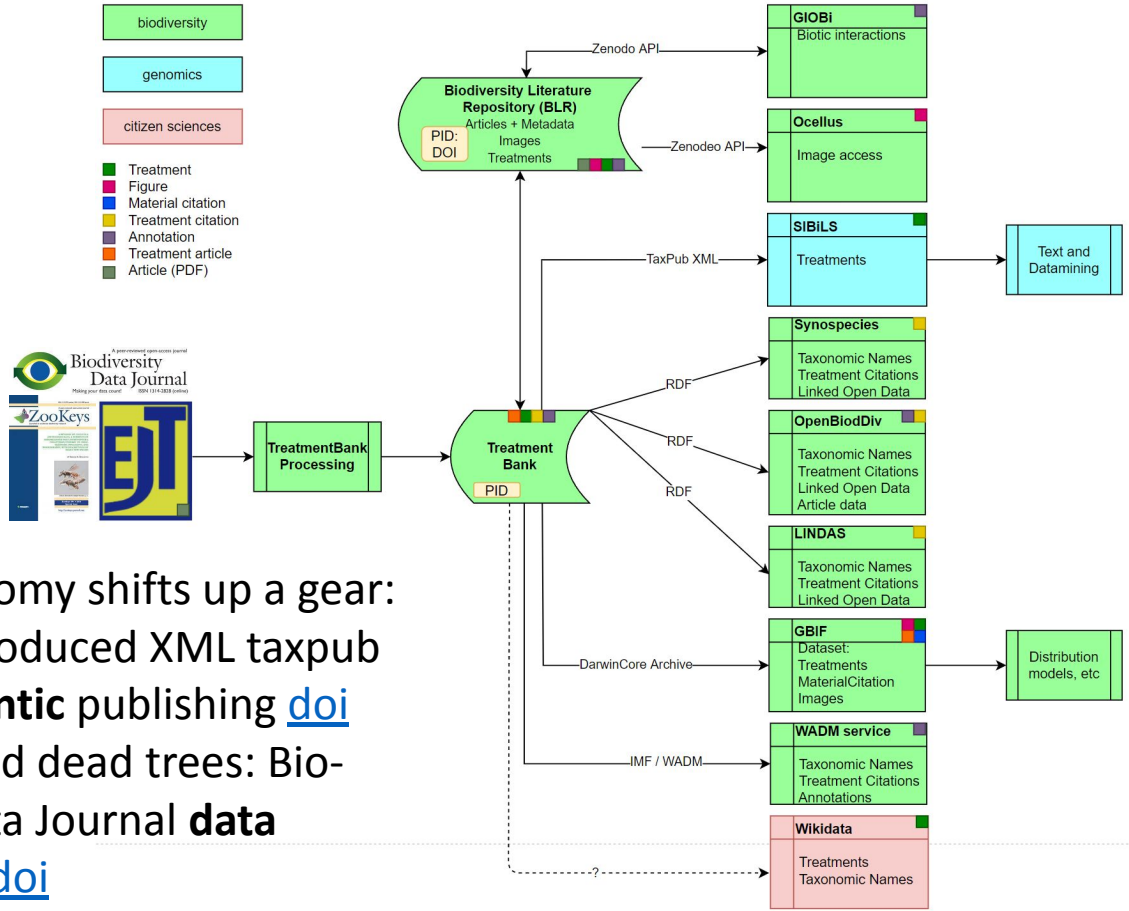
Before antbase.org, Harvard‘s Museum of Comparative Zoology could claim to be the only location with a complete set of ant systematics publications from 1758 - present.





Through antbase.org‘s digital library, access to this body of literature is worldwide, and it is actively used (>10,000 visits in one month only).

2022: Immediate reuse of published data



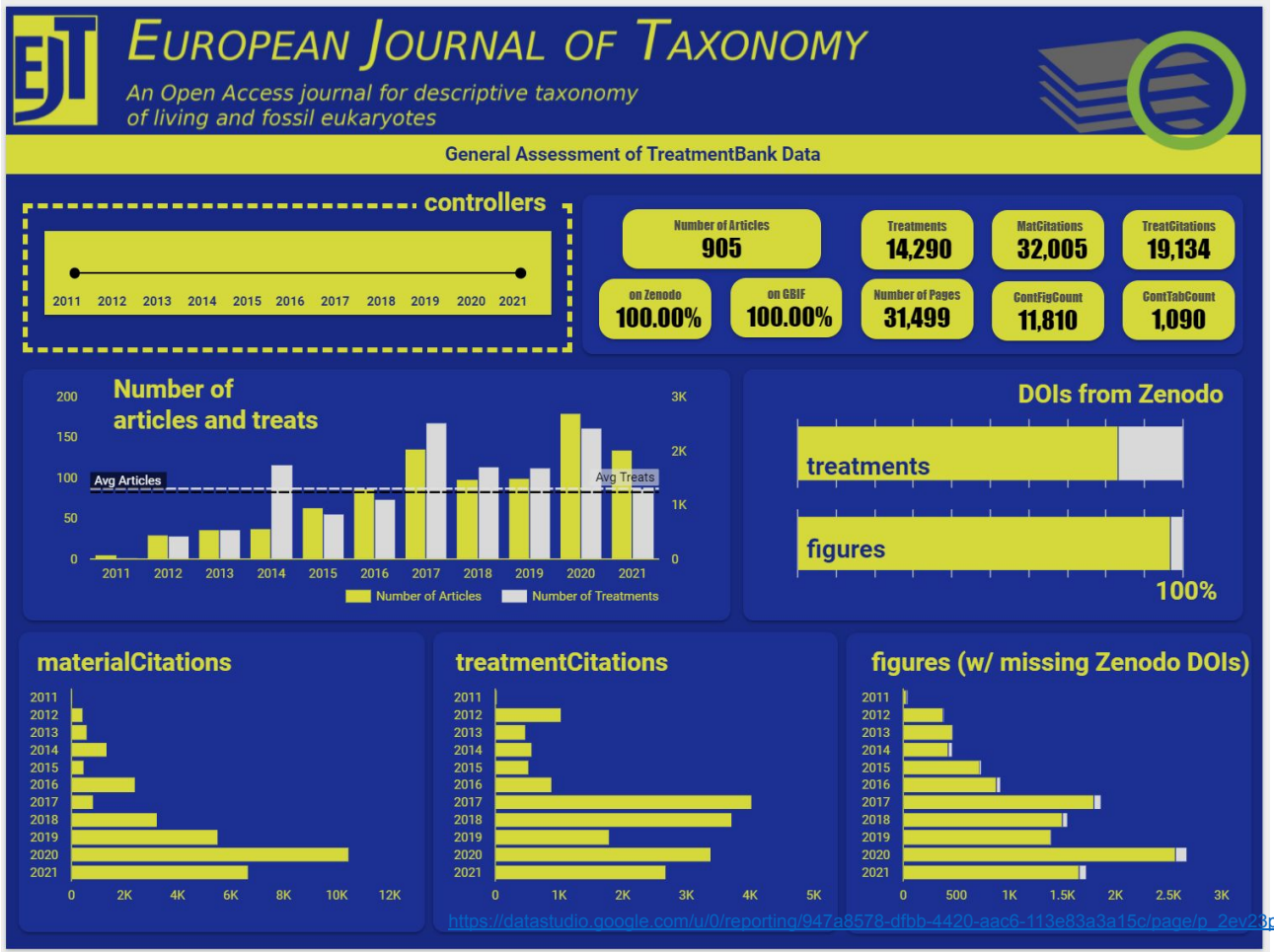
2010: Taxonomy shifts up a gear: Zookeys introduced XML taxpub based **semantic** publishing doi
2013: Beyond dead trees: Bio-diversity Data Journal **data publishing**. doi

BUT: we still do not know what we know.

Publishing specimen data: Material citation: access to data about a specimen (occurrence)

GBIF species: https://www.gbif.org/species/1703942
GBIF occurrence: https://www.gbif.org/occurrence/3015409563
BLR Taxonomic treatment: https://doi.org/10.5281/zenodo.4397319
BLR Article: https://zenodo.org/record/4397315
PDF: https://doi.org/10.11646/zootaxa.4820.1.4

Data analytics



https://datastudio.google.com/u/0/reporting/947a8578-dfbb-4420-aac6-113e83a3a15c/page/p_2ev23pppnc?s=g0qfnzavdgs

What do I know about one treatment?

The use of treatment citation

LESTES (LESTES) BARBARUS (Fabricius) Text fig. 11, 21, 22, pp. 347, 366

Agrion barbara Fabricius, 1798: 286; van der Linden, 1825: 36

Lestes barbara Hagen, 1849: 147 (type species); Hagen & Selys, 1850: 161; Selys, 1862: 318 (34 sep.); Hagen, 1863: 194; McLachlan, 1889: 348; Martin, 1910: 82; Fraser, 1933: 49 fig. 20

Lestes barbarus Selys, 1887: 67 (Egypt); Roster, 1888: 164 pl. 3 (early stages); Kirby, 1890: 162; Martin, 1910: 84, 87; Morton, 1924: 30; May, 1933: 31, 89 (adult and larva); Cowley, 1940: 174; Schmidt, 1951: 122 (quoting Hagen, 1849, *Lestes barbarus* as type species of *Lestes*); Conci & Nielsen, 1956: 65, 66 figs.; Aguesse & Pruja, 1958: 103 (Morocco); Robert, 1958: 82; Loibl, 1958: 55-80; Corbet, 1962: 178; Lieftinck, 1966: 10 (Morocco)

Agrion nympha Hansemann, 1823: 161 (syn. Kirby, 1890)

Lestes barbata Belyshev, 1973: 509 fig. 213 (? laps. cal.)

Type from Barbary in North Africa.

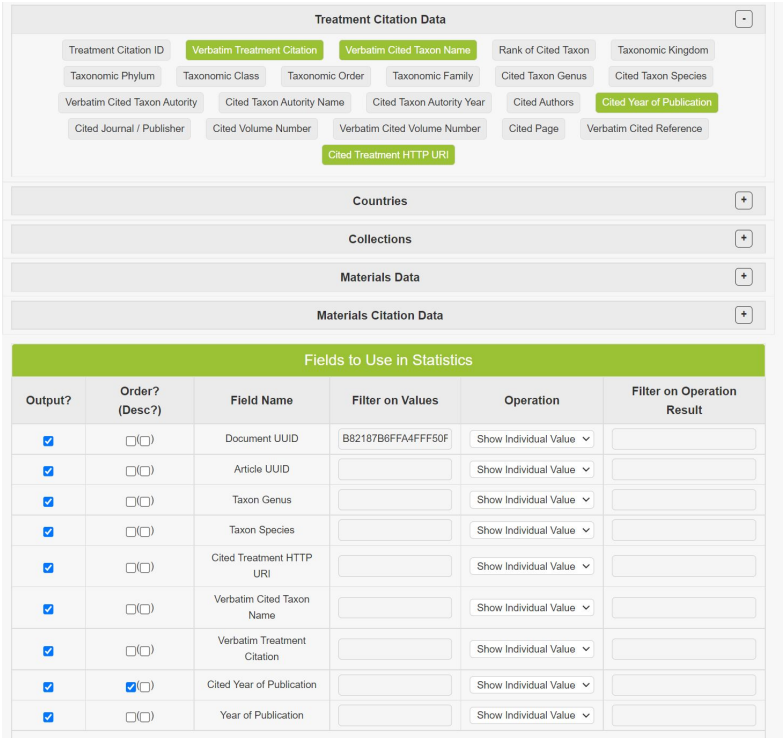
Martin (1910) says (p. 82) "Enfin, à juger d'après un exemplaire rapporté de l'Afrique tropicale par M. Ch. Alluaud, *Lestes barbara*, l'espèce européenne, habite également l'Afrique chaude". On p. 87, in placing *Lestes barbarus* near *obscurus* and *niger* he gave the specific distribution as central and southern Europe, Asia Minor and from there to India; throughout Mediterranean Africa; and states that Alluaud collected the only known record for tropical Africa. He said it was identical to the European form. It would seem probable that Alluaud's *barbara* was either a misidentification, or the locality was incorrect.

Characters. Non pruinose. Pterostigma bicolorous. Wing apices broadly rounded, as in *dryas* and *ictericus*. Sectors of arculus well forward as in *sponsa*, *viridis*, and all Ethiopian species. Orbits below and sternites unmarked with black. Anal appendages yellow with black apex; inferior appendage over half as long as superior, with flat extension ending in an out-turned point. Vulvar scale simple.

General description. Non-pruinose.

Almost mature male (*Morocco*). Labrum greenish yellow. Postclypeus bronze-black, yellow at lateral margins. Head above bronze or metallic green, but antennal bases yellow. Orbits below yellow.

doi



Article stats

Treatment stats

output

PLAZI
PENSOFT
zenodo
CERN
ARCADIA
A CHARITABLE FUND OF LISBET RAUSING & PETER BALDWIN

data liberation infrastructure

Get publications:
- PDF (library access)
- Zenodo via microservice

Receive publications:
- PDF
  - Born digital
  - Scanned
- XML
  - JATS
  - TaxPub/JATS
- HTML / XHTML

PDF

PENSOFT

Decode
- manual
- automated

Enhance
- semantics
- link

Data Quality Control
- Criteria
- Corrections

Create
- Open FAIR data

FAIR Digital Objects

TreatmentBank

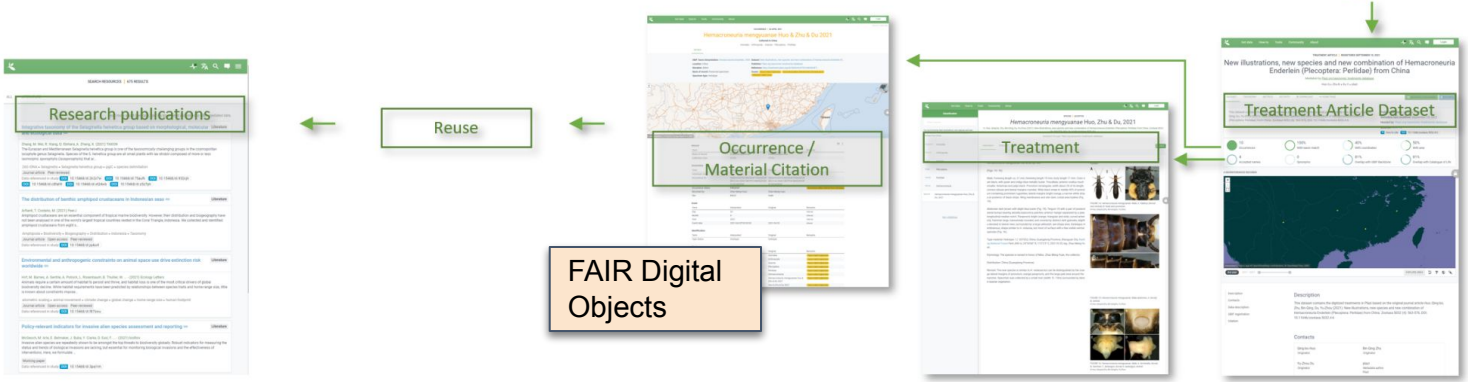
Access / Metrics
- User Interface
- API
- Dashboards

SynoSpecies
Taxonomic names
Treatments

Disseminate
- GBIF
- NCBI

Reuse

GBIF



FAIR Digital Objects

Plazi - Zenodo - GBIF

An example of fully automated processing
Life demonstration to the Flemish government, September 13, 2021

- Original article: https://doi.org/10.11646/zootaxa.5032.4.6
- **Zenodo** article deposition: https://zenodo.org/record/5499090#.YTtdHJ0zb8A
- Zenodo figure deposition: https://zenodo.org/record/5500033#.YTtdMJ0zb8A
- Zenodo/BLR treatment deposition: https://doi.org/10.5281/zenodo.5500037 (see the custom metadata)
- **TreatmentBank** article: https://treatment.plazi.org/GgServer/summary/C154DB54FFA9B65D2460FFED5F578933
- TreatmentBank treatment (HTML): http://treatment.plazi.org/id/3D6DA32C-FFA1-B655-24F7-FDAA5C4F8EC3
- TreatmentBank treatment (JSON): https://zenodo.org/record/5500037/export/json
- TreatmentBank stats: https://tb.plazi.org/GgServer/dioStats/stats?outputFields=doc.articleUuid+doc.doi+doc.zooBankId+doc.gbifId+doc.zenodoDepId+bib.source+cont.pageCount+cont.treatCount+cont.treatCountDoi+cont.treatCitCount+cont.matCitCount+cont.figCount+cont.figCountZen+cont.bibRefCount&groupingFields=doc.articleUuid+doc.doi+doc.zooBankId+doc.gbifId+doc.zenodoDepId+bib.source&FP-doc.articleUuid=C154DB54FFA9B65D2460FFED5F578933&format=JSON
- **GBIF** dataset ID: https://www.gbif.org/dataset/8f239084-30f3-4a6c-ba97-3eb65356beb5
- GBIF occurrence data set: https://www.gbif.org/occurrence/search?dataset_key=8f239084-30f3-4a6c-ba97-3eb65356beb5
- GBIF occurrence data set, holotypes only: https://www.gbif.org/occurrence/search?dataset_key=8f239084-30f3-4a6c-ba97-3eb65356beb5&type_status=HOLOTYPE
- GBIF species pages: https://www.gbif.org/species/search?dataset_key=8f239084-30f3-4a6c-ba97-3eb65356beb5&origin=SOURCE&status=ACCEPTED&advanced=1
- **Ocellus** (Bilder in BLR): https://ocellus.info/images.html?q=%2210.11646/zootaxa.5032.4.6%22&size=30&page=1&communities=biosyslit
- **Synospecies** (Synoynym triple store): https://synospecies.plazi.org/#Kiotina+spatulata

Infrastructures maintained by Plazi

Biodiversity Literature Repository

Repository for data liberated from publications
453,000 images
72,000 articles
400,000 taxonomic treatments
Collaboration with Zenodo /CERN
Recognized as EU research infrastructure
Mints DOI for treatments and figures

TreatmentBank

Data conversion and access service
73,000 articles
762,000 taxonomic treatments
1,118,000 materials citations
>50% of annually described new species
production in 2021: 25,000 articles, 224,000 treatments, 180,000 images

Plazi mediated data in GBIF

Global Biodiversity Information Facility

Reuse of treatment articles mediated by Plazi

- 39,700 treatment article data sets (58% of total data sets in GBIF)
- 377,000 taxonomic treatments (90,000 unique species)
- 224,000 figures
- 591,000 materials citations (occurrences)
  - The figures are lower in GBIF because of Quality Control measures

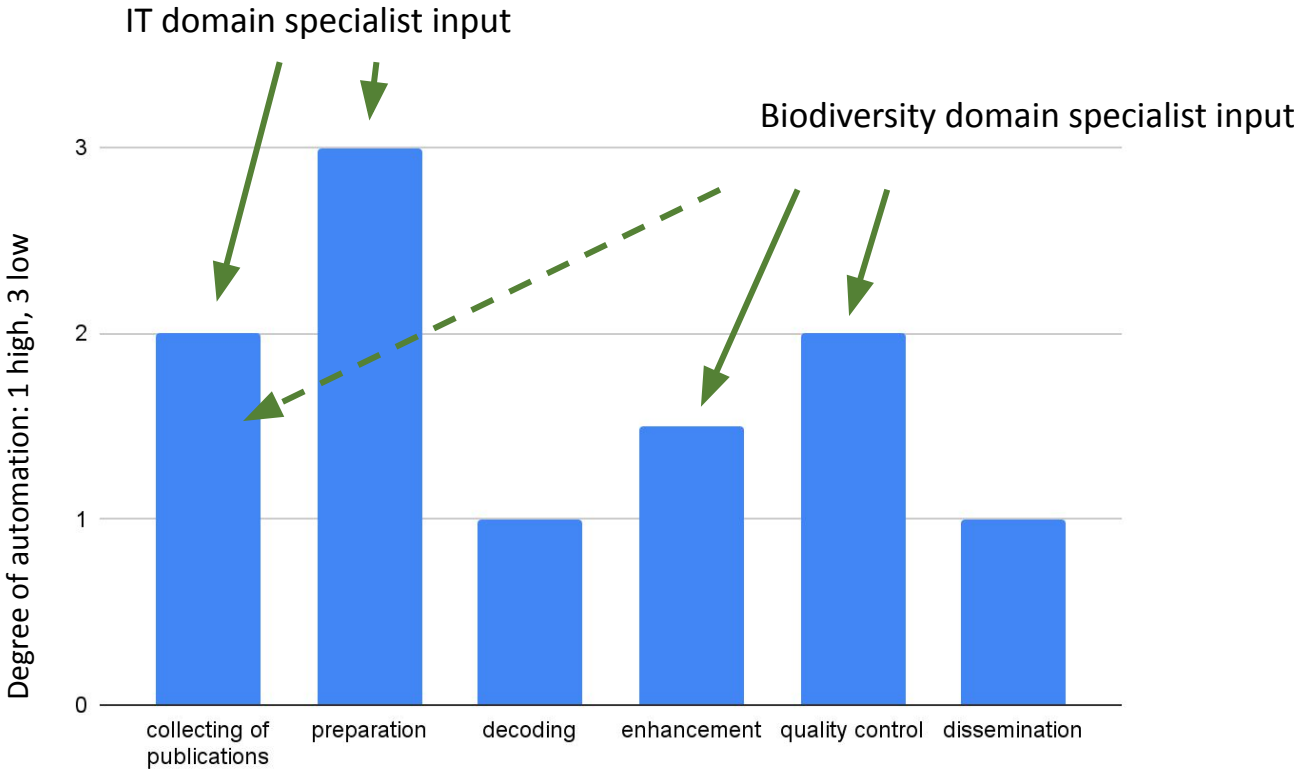
In comparison to what's in TreatmentBank

Data conversion and access service

- 73,000 articles
- 762,000 taxonomic treatments
- 1,118,000 materials citations
  - >50% of annually described new species
  - production in 2021: 25,000 articles, 224,000 treatments, 180,000 images

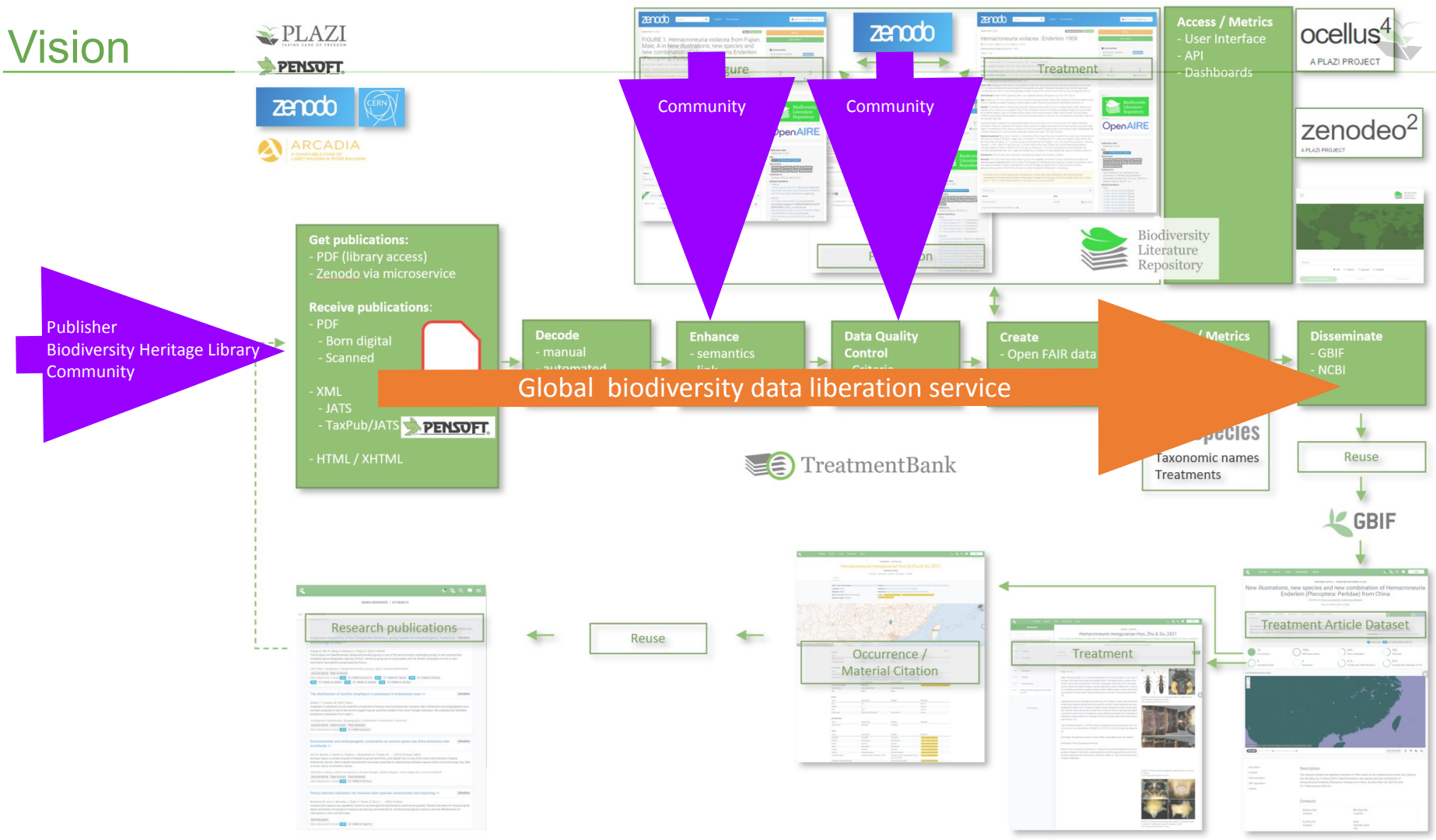
The lower number in GBIF is due to quality control and minimal standards for data targeted for GBIF.

Next steps: getting the crowd involved

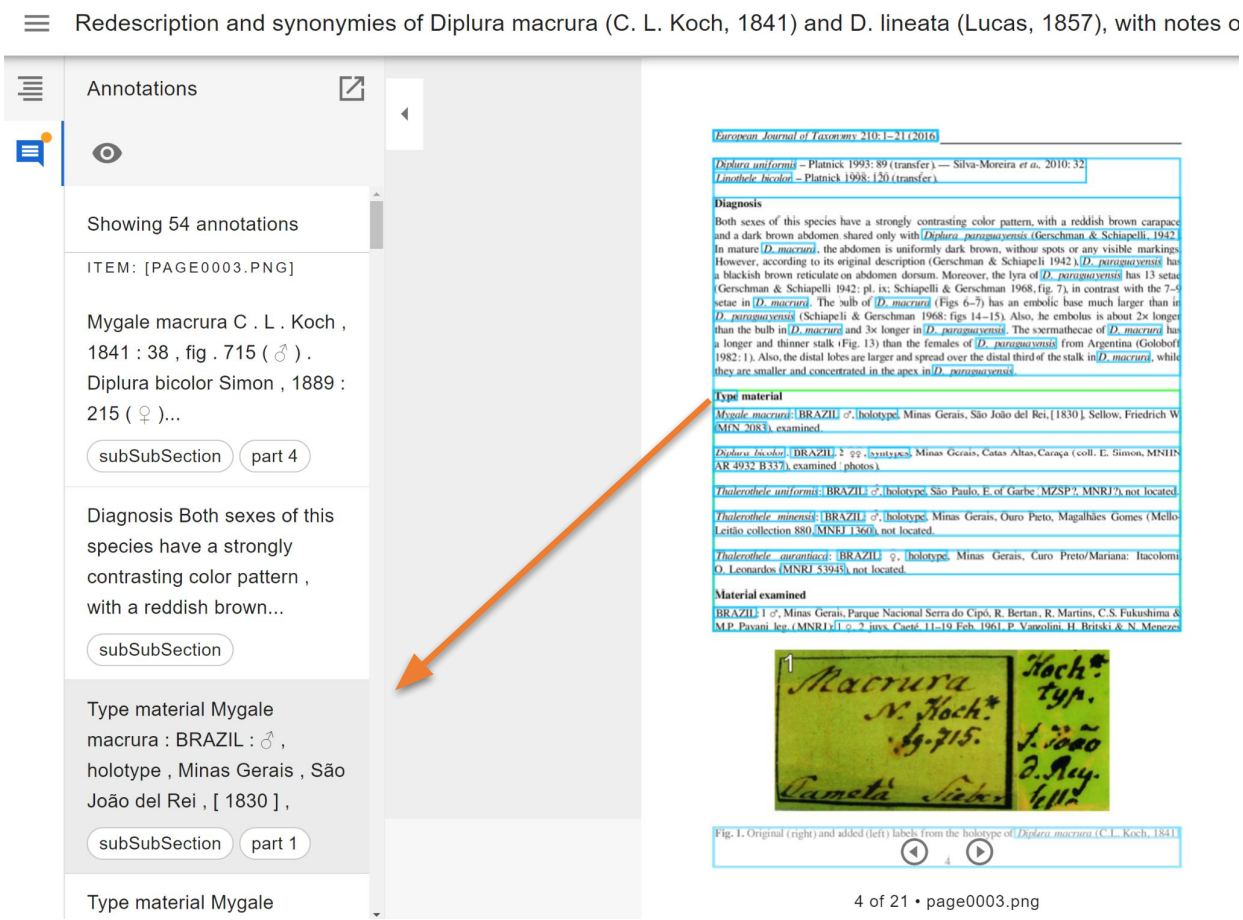


This is a very expensive effort, has to be shared by partners avoiding duplications as much as possible, and needs be avoided by changing the way we publish.

Vision


Community
Community
Access / Metrics
- User Interface
- API
- Dashboards
Treatment
Biodiversity Literature Repository
Publisher
Biodiversity Heritage Library
Community
Get publications:
- PDF (library access)
- Zenodo via microservice
Receive publications:
- PDF
- Born digital
- Scanned
- XML
- JATS
- TaxPub/JATS
- HTML / XHTML
Decode
- manual
Enhance
- semantics
Data Quality Control
Create
- Open FAIR data
Disseminate
- GBIF
- NCBI
Global biodiversity data liberation service
TreatmentBank
Taxonomic names
Treatments
Reuse
GBIF
Treatment Article Dataset
Treatment
Occurrence / Material Citation
Reuse
Research publications

Next steps: annotations by the crowd



Source

Plazi: goal

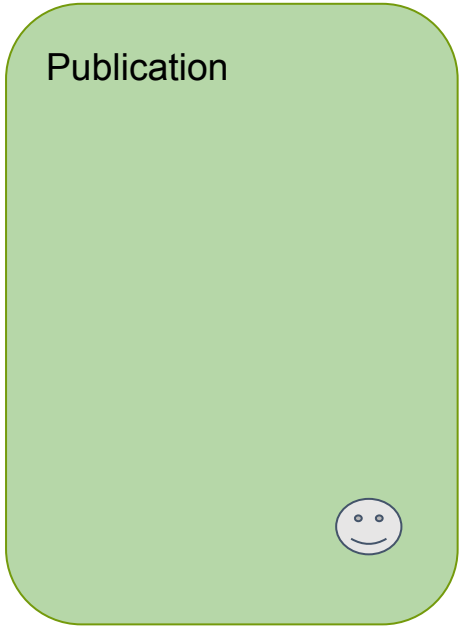
Discovering known biodiversity

Create a list of the Earth' known taxa, and related digital accessible knowledge embedded in scholarly publications as open findable, accessible, interoperable and reusable data about the Earth's species (FAIR digital objects), as input to the biodiversity knowledge graph, liberated from scholarly publications.

Digital accessible knowledge in biodiversity (DAK)

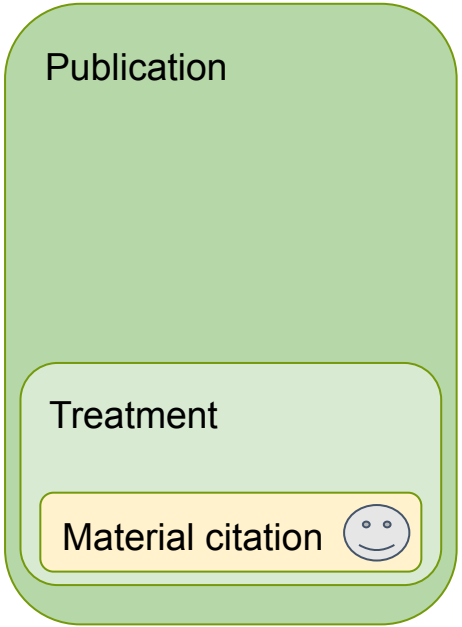
- Data understandable by human and actionable by machine reflecting the growth of our knowledge and interdependence.
- Editorial structure including textflow, paragraphs, sections such as title, authors and affiliations, materials and methods, etc., tables, figures, bibliographic references and their citations in the text.
- Semantic annotated (e.g. using TDWG standards) data at the base of the biodiversity knowledge graph:
    - data about a taxon (taxonomic treatments with their nomenclature section)
    - cited previous treatments (treatment citations)
    - cited specimens (material citations)
    - named entities (persons, taxonomic names, accession -, collection -, institution - or specimen codes
    - attributes including their persistent identifiers

Annotations are context, defining the role of named entities



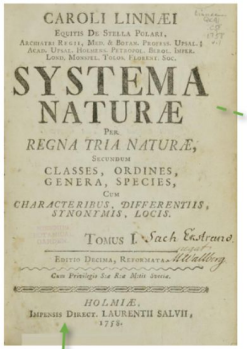
Person: **role**
Public. author, authority, collector

Person: **role**
Publication
Treatment authority, collector



Person: **role**
Publication
Treatment
Material citation collector

Digital Accessible Knowledge: Treatment


CAROLI LINNÆI
EQUITIS DE STELLA POLARI,
ARCHIATRI REGII, MED. & BOTAN. PROFESS. UPSAL.
ACAD. UPSAL. HOLMENS., PETROPOL., BERAN., IMPER.
LOND., MONSPEL., TOLOS., FLORENT. SOC.

SYSTEMA NATURÆ

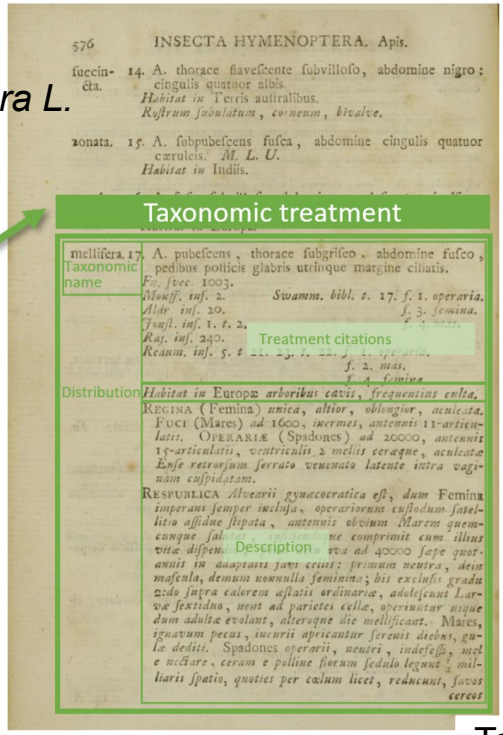
PER REGNA TRIA NATURÆ,
SECUNDUM CLASSES, ORDINES, GENERA, SPECIES,
CUM CHARACTERIBUS, DIFFERENTIIS, SYNONYMIS, LOCIS.

TOMUS I. Sach Exstrans.

HOLMIÆ,
IMPENSIS DIRECT. LAURENTII SALVII,
1758.

Linnaeus, 1758
4,819 taxa

Apis mellifera L.


576 INSECTA HYMENOPTERA. Apis.

succinc-eta. 14. A. thorace flavescente subvilloso, abdomine nigro; cingulis quatuor albis.
Habitat in Terris australibus.
Rostrum subulatum, corneum, bivalve.

zonata. 15. A. subpubescens fusca, abdomine cingulis quatuor cæruleis. *M. L. U.*
Habitat in Indiis.

Taxonomic treatment

mellifera. 17. Taxonomic name A. pubescens, thorace subgriseo, abdomine fusco, pedibus posticis glabris utrinque margine ciliatis.
Fn. suec. 1003.
Mouff. inf. 2. *Swamm. bibl. t.* 17. *f.* 1. *operaria.*
Aldr. inf. 20. *f.* 3. *femina.*
Jonst. inf. t. t. 2.
Raj. inf. 240.
Reaum. inf. 5. *f.*
f. 2. *mas.*
f. 4. *femina.*

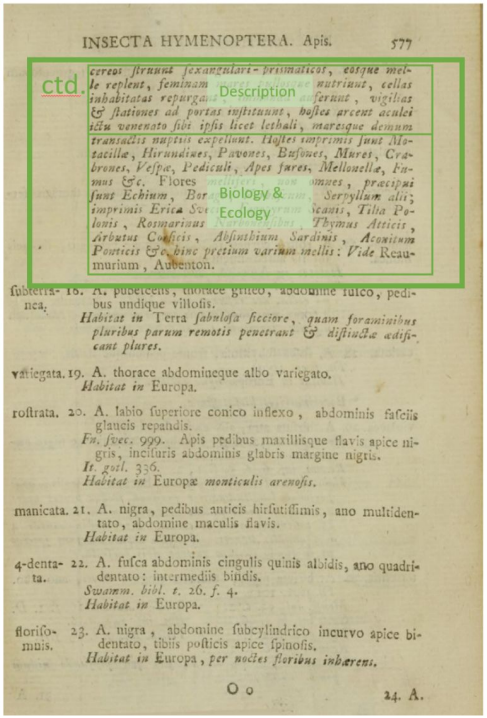
Treatment citations

Distribution *Habitat in* Europæ *arboribus cavis, frequentius exilta,*

REGINA (Femina) *unica, altior, oblongior, aculeata.* FUCI (Mares) *ad* 1600, *inermes, antennis* 11-*articulatis.* OPERARIÆ (Spadones) *ad* 20000, *antennis* 13-*articulatis, ventriculis* 2 *melliferæ, aculeatæ Ense retrorsum ferrato venenato latente intra vaginam cuspidatam.*

RESPUBLICA *Alvearii gynæocratica est, dum* Femina *imperans semper inclusa, operariorum custodum satellitio assidue stipata, antennis obvium Marem quemcunque salutantibus comprimit cum illius vitæ dispendio. Ova ad* 40000 *sæpe quotannis in duabus tabulis cerei: primum neutra, dein masculæ, demum nonnulla femininæ; his exclusis gradu calido supra calorem æstatis ordinariæ, adolescunt Larvæ fœtidæ, nent ad parietes cellæ, operiuntur usque dum adultæ evolant, alteroque die mellificant. Mares, ignavum pecus, incurii apricantur serenis diebus, gulæ dediti. Spadones operarii, neutri, indefessi, mel e necture, ceram e polline florum sedulo legunt; milliaris spatio, quoties per cælum licet, redeunt, favos cereos*

Description


INSECTA HYMENOPTERA. Apis. 577

ctd. *cereos struunt hexangulari-prismaticos, eosque mellæ replent, seminam nutriunt, cellas inhabitatas repurgant, feruntque, vigilias & stationes ad portas instituunt, hostes arcent aculei ictu venenato sibi ipsis licet lethali, mareique demum transactis nuptiis expellunt. Hostes imprimis sunt Motacillæ, Hirundines, Pavones, Buffones, Mures, Crabrones, Vespæ, Pediculi, Apes furesque, Melloucllæ, Eumus &c. Flores omnes, præcipui sunt Echium, Bor Serpyllum alii; imprimis Erica Sco Scabiosa, Tilia Polonis, Rosmarinus Thymus Atticis, Arbutus Cretensis, Absinthium Sardinis, Aconitum Ponticis &c. hinc pretium varium mellis: Vide Reaumurium, Aubenton.*

Description

Biology & Ecology

subcris- 18. A. pubescens, thorace griseo, abdomine fusco, pedinea. bus undique villosis.
Habitat in Terra *fabulosa sicciore, quam foraminibus pluribus parum remotis penetrant & distinctæ ædificant plures.*

variegata. 19. A. thorace abdomineque albo variegato.
Habitat in Europa.

rostrata. 20. A. labio superiore conico inflexo, abdominis fasciis glaucis repandis.
Fn. suec. 999. Apis pedibus maxillisque flavis apice nigris, incisuris abdominis glabris margine nigris.
It. gotl. 336.
Habitat in Europæ *monticulis arenosis.*

manicata. 21. A. nigra, pedibus anticis hirsutissimis, ano multidentato, abdomine maculis flavis.
Habitat in Europa.

4-denta- 22. A. fusca abdominis cingulis quinis albidis, ano quadri-ta. dentato: intermediis bindis.
Swamm. bibl. t. 26. *f.* 4.
Habitat in Europa.

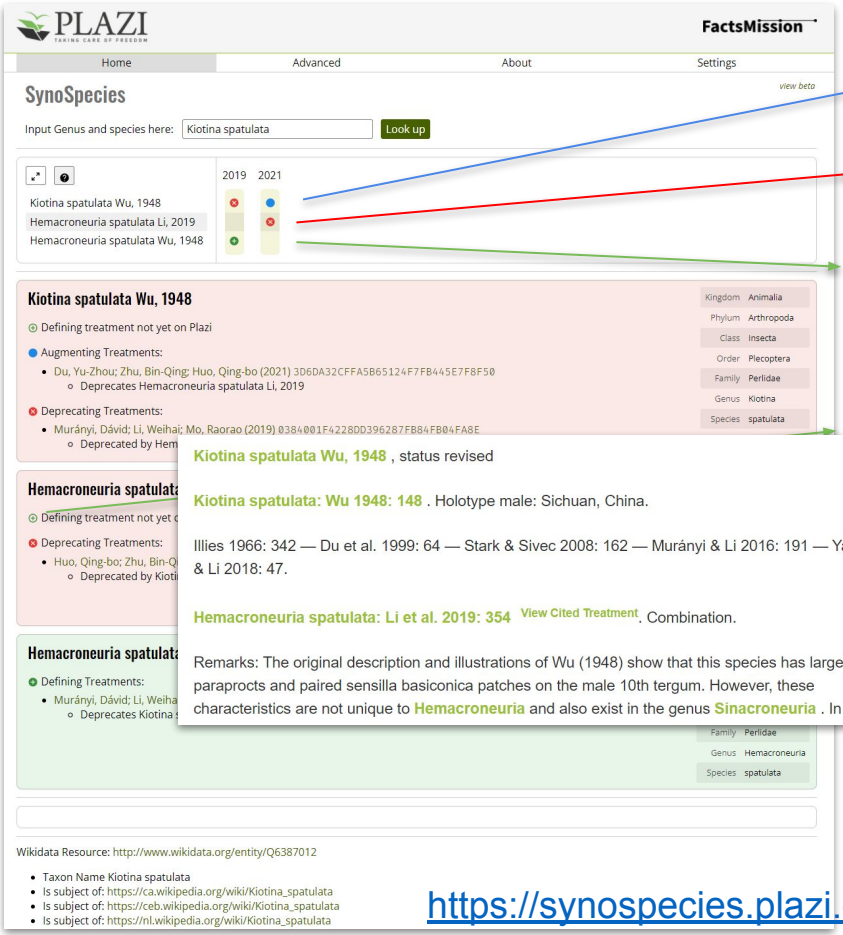
florisso- 23. A. nigra, abdomine subcylindrico incurvo apice bi-mus. dentato, tibiis posticis apice spinosis.
Habitat in Europa, *per noctes floribus inhærens.*

O o 24. A.

Each type has a published taxonomic treatment
Each taxonomic treatment is multiple times augmented

Tens of millions exist as part of ca 500 Million published pages of biodiversity literature
Each includes a plethora of facts

Treatment citation: Catalogue of life



Currently accepted name

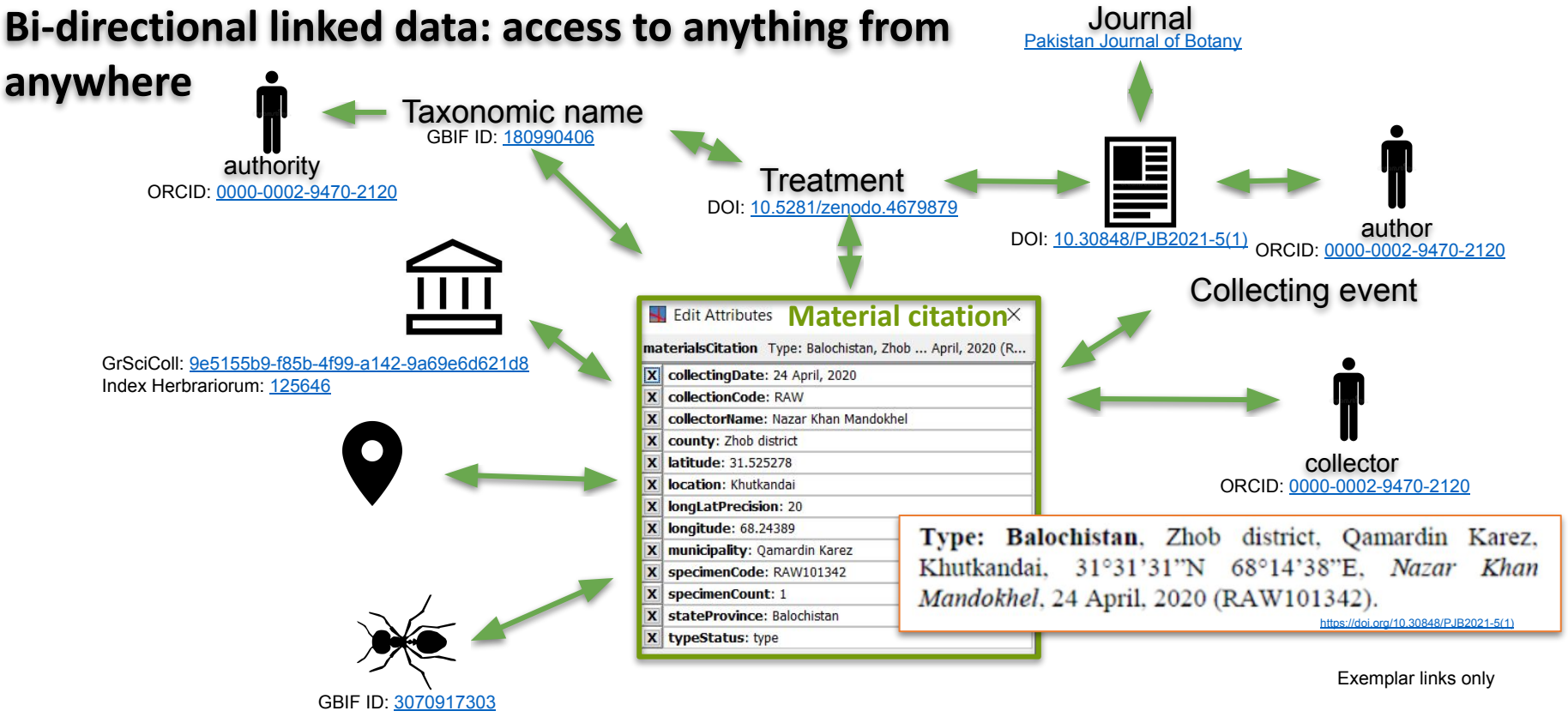
Synonymized taxonomic name

Original name or new combination

Missing treatment

https://synospecies.plazi.org/#Kiotina+spatulata

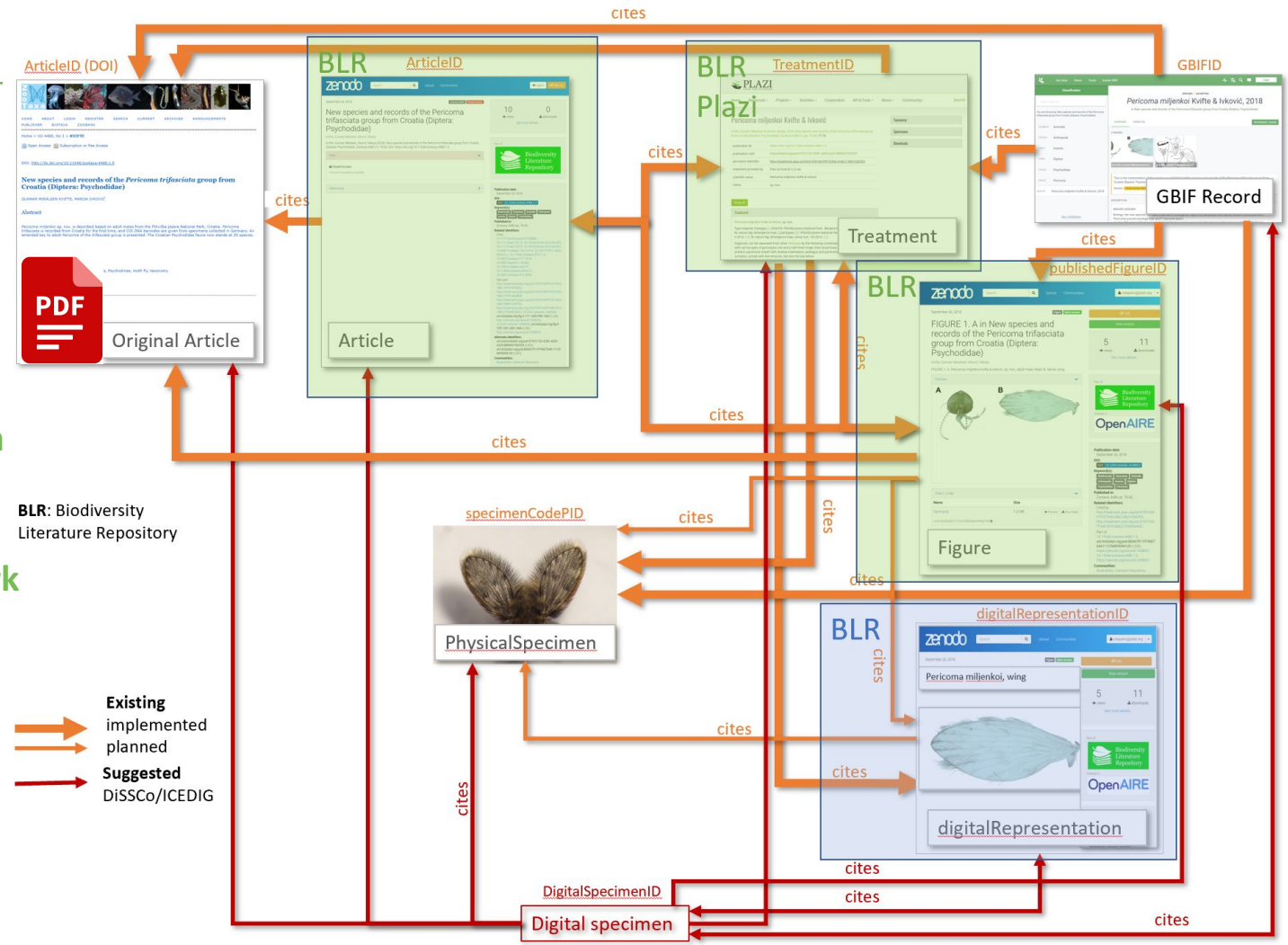
Published materials citation links



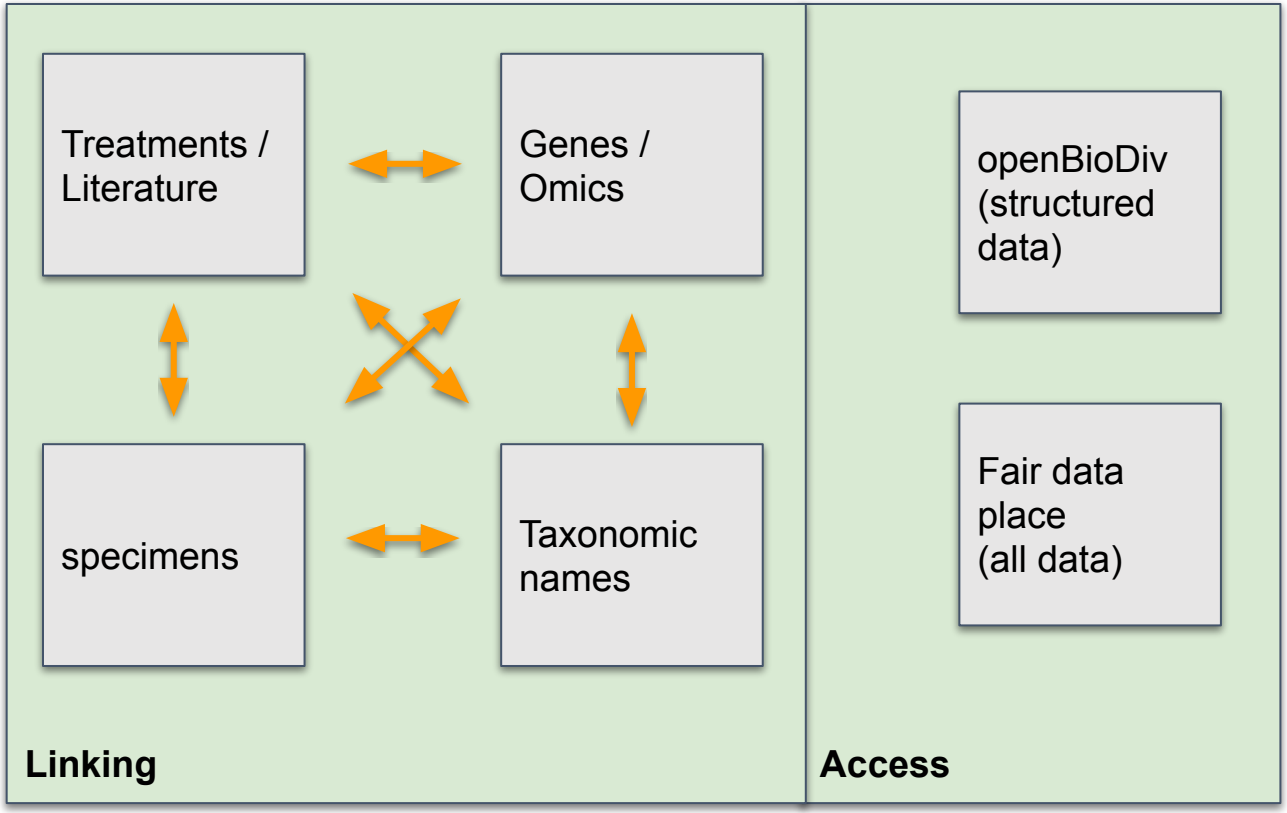
Imagine the possible applications enabling making use of this big data?
Imagine the time saved if all these links are hyperlinks?

Focus on data in publications: What's inside?

Citation network



BiCIKL: Linking and access

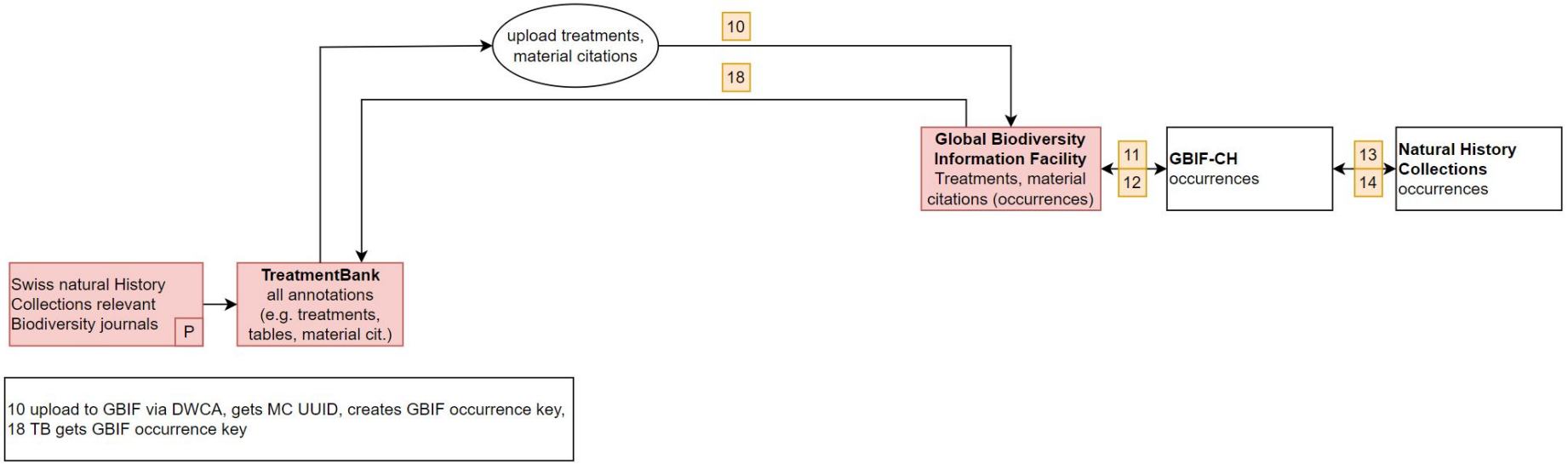


Linking specimens and literature via material citations

Bidirectional linking

material citations

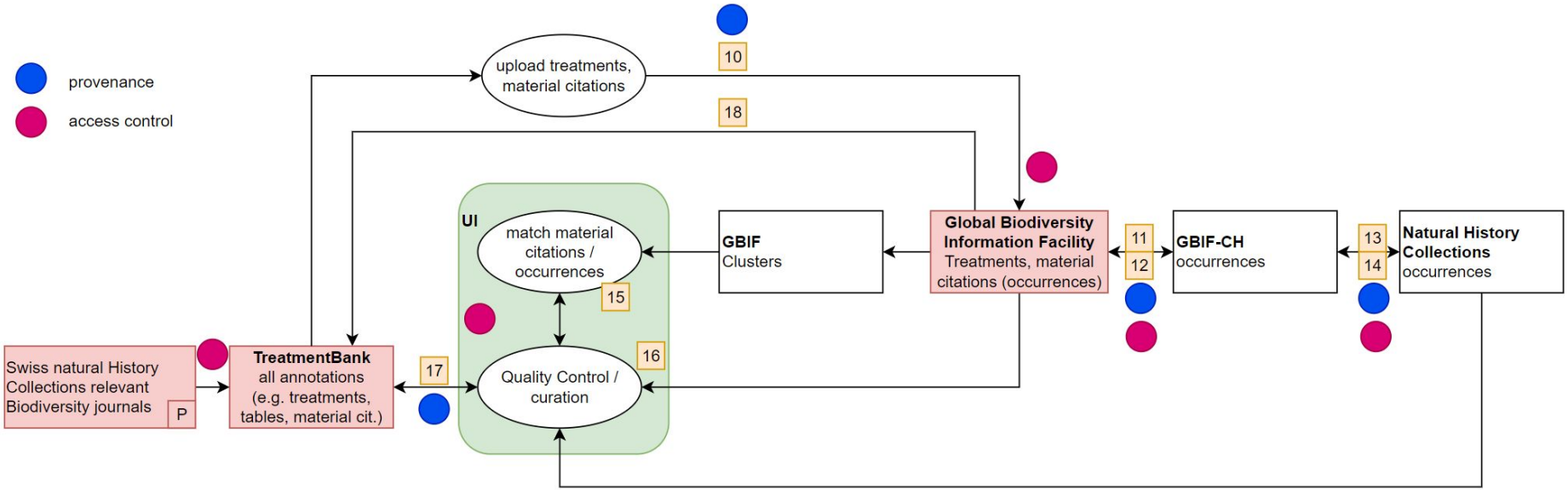
specimens / occurrences



eBioDiv

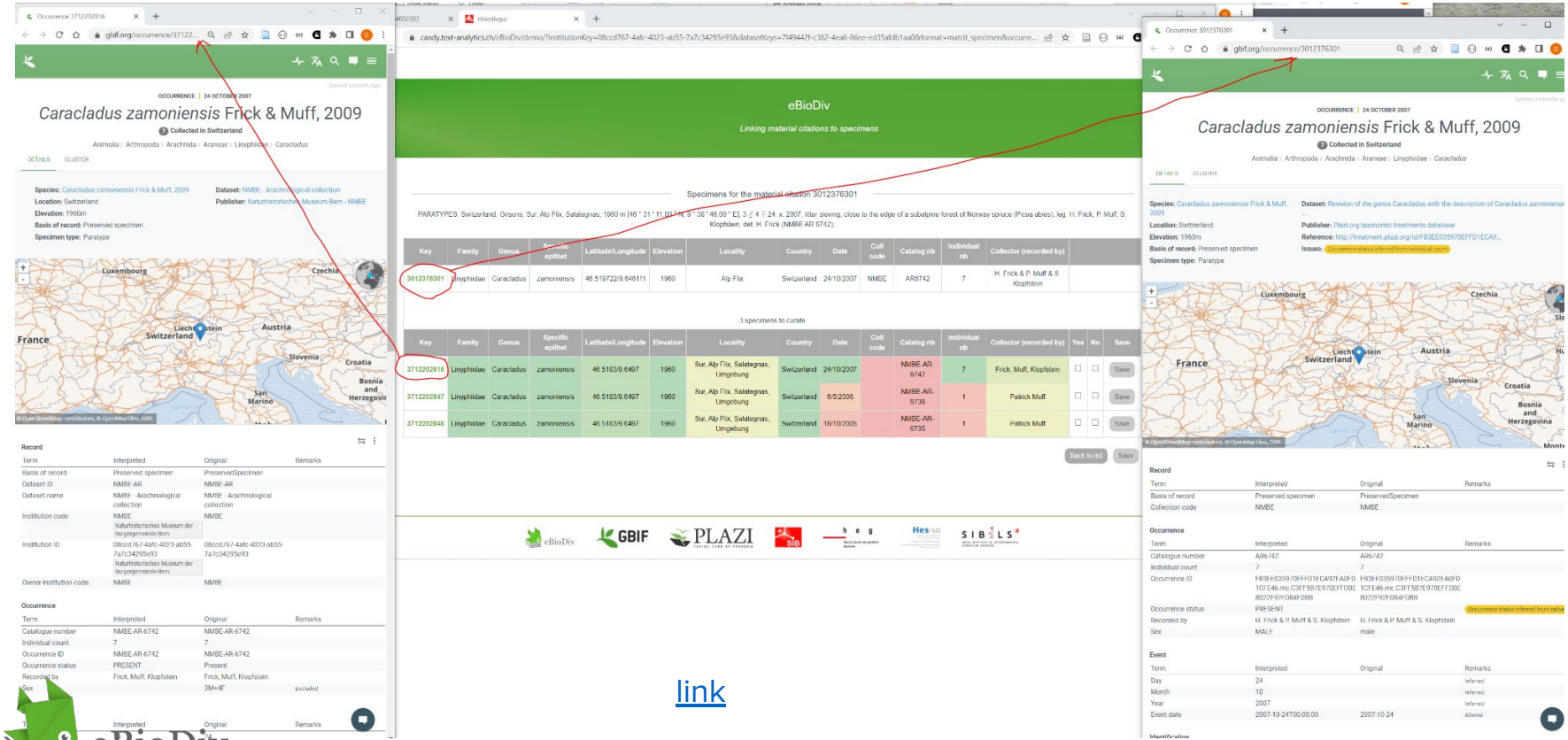
Pilot matching service (link)

Linking specimens and literature via material citations



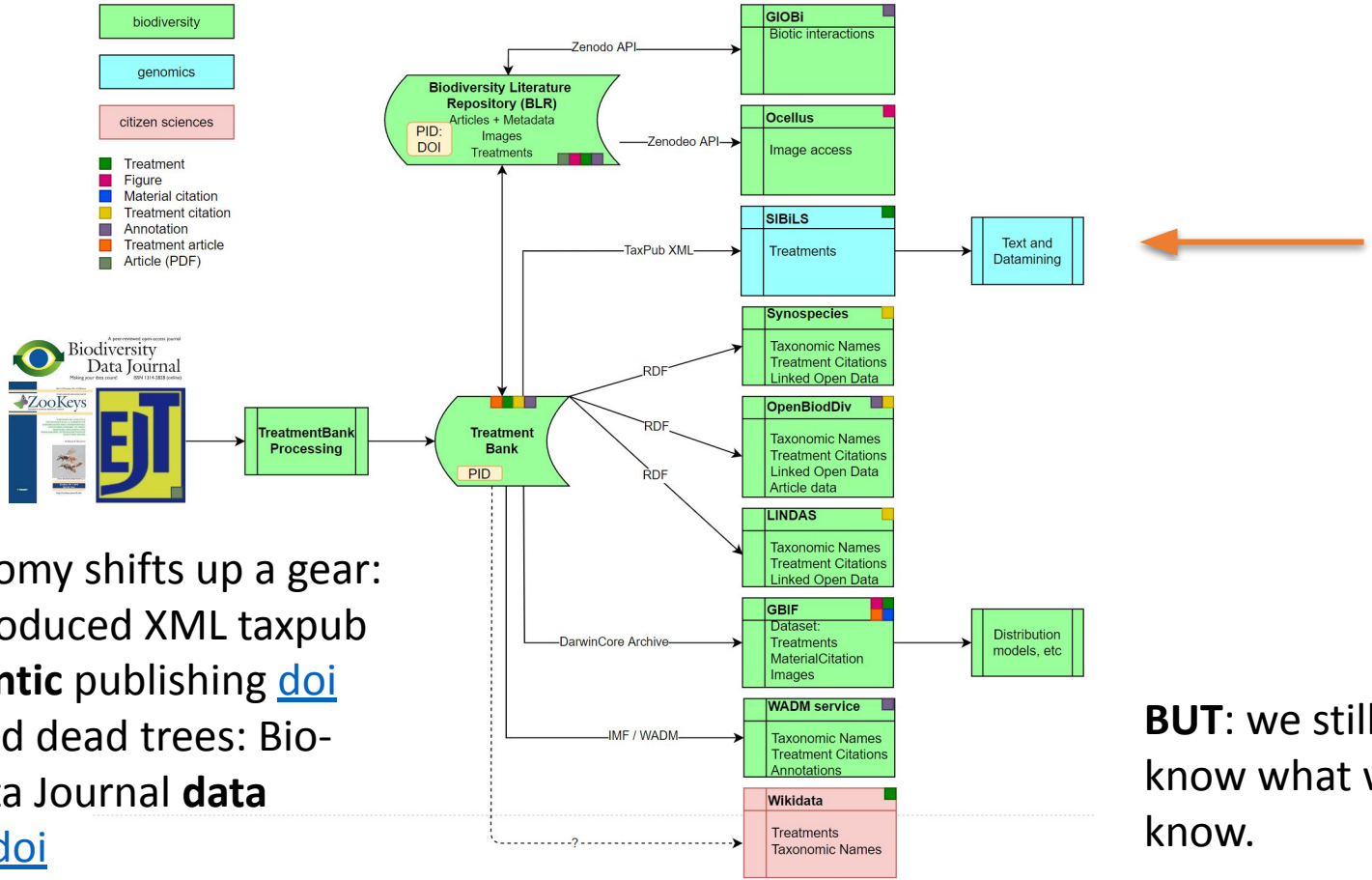
eBioDiv

Linking specimens with material citations



eBioDiv

2022: Building a bridge between biodiversity and bioinformatics



2010: Taxonomy shifts up a gear: Zookeys introduced XML taxpub based **semantic** publishing doi

2013: Beyond dead trees: Bio-diversity Data Journal **data publishing**. doi

BUT: we still do not know what we know.

Bridging domains: Get treatments into the bioinformatics world

Table JSON

SIBiLS_upload_time	May 4, 2022 @ 16:22:54.560
t _id	http://treatment.plazi.org/id/011D87C1FFC1CD711FBFFB9DE3D5FE4D
t _index	baseline_alpha
# _score	1
t _type	_doc
t article-title	Observations on non-didemnid ascidians from Australian waters (1)
t material_citation_distribution	Previously recorded (see Kott 1985): Western Australia (Cape Jaubert, Shark Bay, Cockburn Sound); South Australia (Upper Spencer Gulf); Vic Mile Beach, Bass Strait, Warnambool); New South Wales (Byron Bay); Queensland (Maroochydore, Hervey Bay, Gladstone, southern Great Barrier il); Indonesia, Singapore, Sri Lanka, Japan. New records: Queensland (17.165 –17.935 ° S, 146.535– 146.8 ° E, 35–68 m)
t nomenclature-taxon-name	Polycarpa procera (Sluiter, 1885)
t publication-doi	http://dx.doi.org/10.1080/00222930600621601
t taxon_name_distribution	Polycarpa chinensis (Tokioka, 1967) P. procera Polycarpa chinensis
t taxon_name_reference_group	Styela procera Sluiter 1885, p 196 Polycarpa procera: Kott 1985, p 196
t text	Styela procera Sluiter 1885, p 196 . Polycarpa procera: Kott 1985, p 196 and synonymy. Distribution Previously recorded (see Kott 1985): Western Australia (Cape Jaubert, Shark Bay, Cockburn Sound); South Australia (Upper Spencer Gulf); Vic Mile Beach, Bass Strait, Warnambool); New South Wales (Byron Bay); Queensland (Maroochydore, Hervey Bay, Gladstone, southern Great Barrier il); Indonesia, Singapore, Sri Lanka, Japan. New records: Queensland (17.165 –17.935 ° S, 146.535– 146.8 ° E, 35–68 m) . Remarks
t text_distribution	Distribution Previously recorded (see Kott 1985): Western Australia (Cape Jaubert, Shark Bay, Cockburn Sound); South Australia (Upper Spencer Gulf); Vic Mile Beach, Bass Strait, Warnambool); New South Wales (Byron Bay); Queensland (Maroochydore, Hervey Bay, Gladstone, southern Great Barrier il); Indonesia, Singapore, Sri Lanka, Japan. New records: Queensland (17.165 –17.935 ° S, 146.535– 146.8 ° E, 35–68 m) . Remarks The species is plentiful in inter-reefal locations in northeastern Queensland and a wide range in the Indo-West Pacific between Japan in th i Lanka and around the southern coast of the Australian continent. The species range supports the view that the Australian continental shel
t text_reference_group	Styela procera Sluiter 1885, p 196 . Polycarpa procera: Kott 1985, p 196 and synonymy.
t treatment-bank-uri	http://treatment.plazi.org/id/011D87C1FFC1CD711FBFFB9DE3D5FE4D
t treatment_title	Polycarpa procera
t zenodo-doi	10.1080/00222930600621601

source

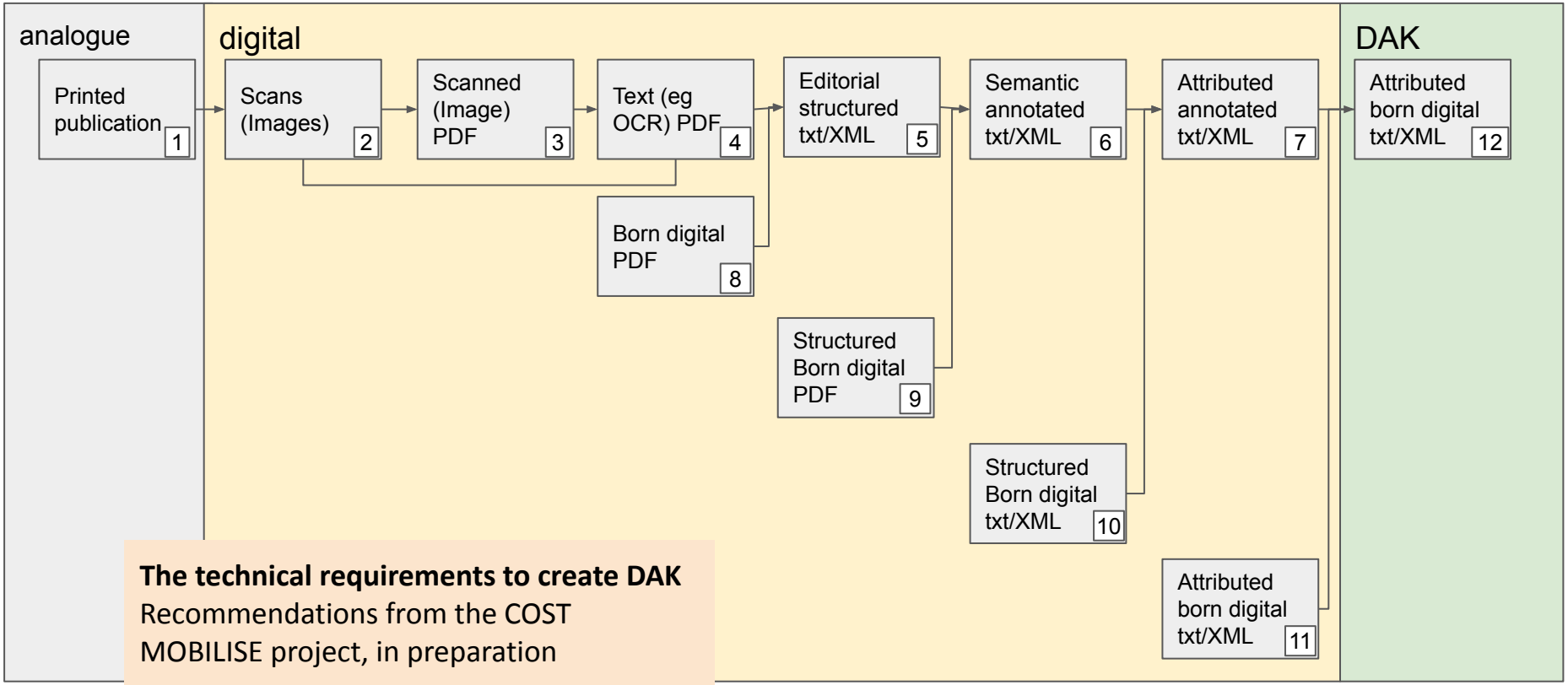
Treatment imported as TaxPub/JATS to Swiss Institute of Bioinformatics Library Service (SIBiLS).

Bioinformatics: Access to biodiversity data

Biodiversity Informatics: Access to tools to annotated and text and data mine treatments (e.g. biotic interactions terms)

eBioDiv

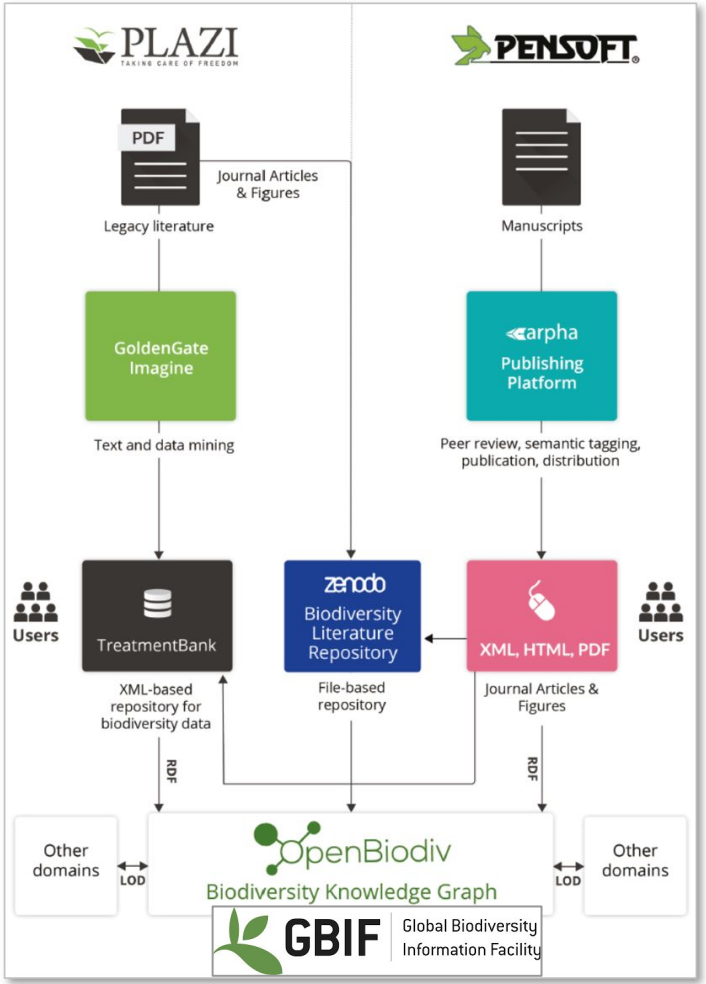
Origin of Digital Accessible Knowledge (DAK)



The technical requirements to create DAK
Recommendations from the COST MOBILISE project, in preparation

The future will be open, semantically enhanced publishing

Legacy publications



Prospective publishing

Thank you!

Questions, answers, participation https://github.com/plazi/community

Introduction to digitizing taxonomic literature with Plazi DOI

Donat Agosti agosti@plazi.org

Appendix

Linking specimens and literature via material citations

Further reading:

- Plazi: further reading
- Pensoft: doi: 10.3897/zookeys.50.543 (e.g. Zookeys, BDJ)
- CETAF: doi: 10.5252/adansonia2018v40a1 (e.g. European Journal of Taxonomy)

Data usage:

- Rivera-Quiroz et al. 2020, doi: 10.1038/s41598-020-72549-8
- Dikow & Agosti, 2015, doi: 10.3897/BDJ.3.e5707

Data access:

- Brief introduction into Treatmentbank stats: PDF
- Treatment statistics: https://tb.plazi.org/GgServer/srsStats
- Article statistics: https://tb.plazi.org/GgServer/dioStats
- Biodiversity Literature Repository API introduction: https://developers.zenodo.org/
- Biodiversity Literature Repository: https://zenodo.org/communities/biosyslit/search?q=

Applications based on and reuse of TreatmentBank and BLR data:

- Images via Ocellus: https://ocellus.info/
- Treatment citations via Synospecies: https://synospecies.plazi.org/
- TreatmentBank data in GBIF: https://www.gbif.org/publisher/7ce8aef0-9e92-11dc-8738-b8a03c50a862

Plazi

Plazi is a Swiss based international association supporting and promoting the development of persistent and openly accessible scholarly digital taxonomic publications

NGO, SME owned by the NGO; Founded in 2008 as spin-off from a former US/DFG binational digital library award (2003-06); Supported by service contracts, EU-research funding, philanthropic funds, voluntary contributions. Plazi GmbH SME as service provider.

13 persons working for Plazi in Brazil, France, Germany, Spain, Switzerland, USA

Collaborations with Global Biodiversity Information Facility (GBIF), Zenodo at CERN, Pensoft Publishers Ltd, Consortium of European Taxonomic Facilities (CETAF), Swiss Institute of Bioinformatics (SIB), National Center for Biotechnology Informatics (NCBI), Muséum nationale d'Histoire Naturelle, Paris, Data Futures.

Support from Arcadia Fund, EU projects, service contracts, and voluntary contributions.

A mission of Plazi is to **discover, make accessible, and disseminate known biodiversity data,** not publications *per se* and to promote semantic enhanced publishing (TaxPub/JATS).