Doi: 10.5281/zenodo.6590938

Nothing in (taxonomic) publishing makes sense except in the light of treatments

June 7, 2022

SPNHC

Edinburgh

Donat Agosti

ORCID: 0000-0001-9286-1200; @myrmoteras

Plazi, Switzerland

SIBILS
SWISS INSTITUTE OF BIOINFORMATICS
LITERATURE SERVICES

swissuniversities

101007492

ARCADIA

BiCIKL

eBioDiv

Biodiversity Literature Repository

Mobilise Action
Mobilising Data, Policies and Experts in Scientific Collections

TreatmentBank

Taxonomic publishing

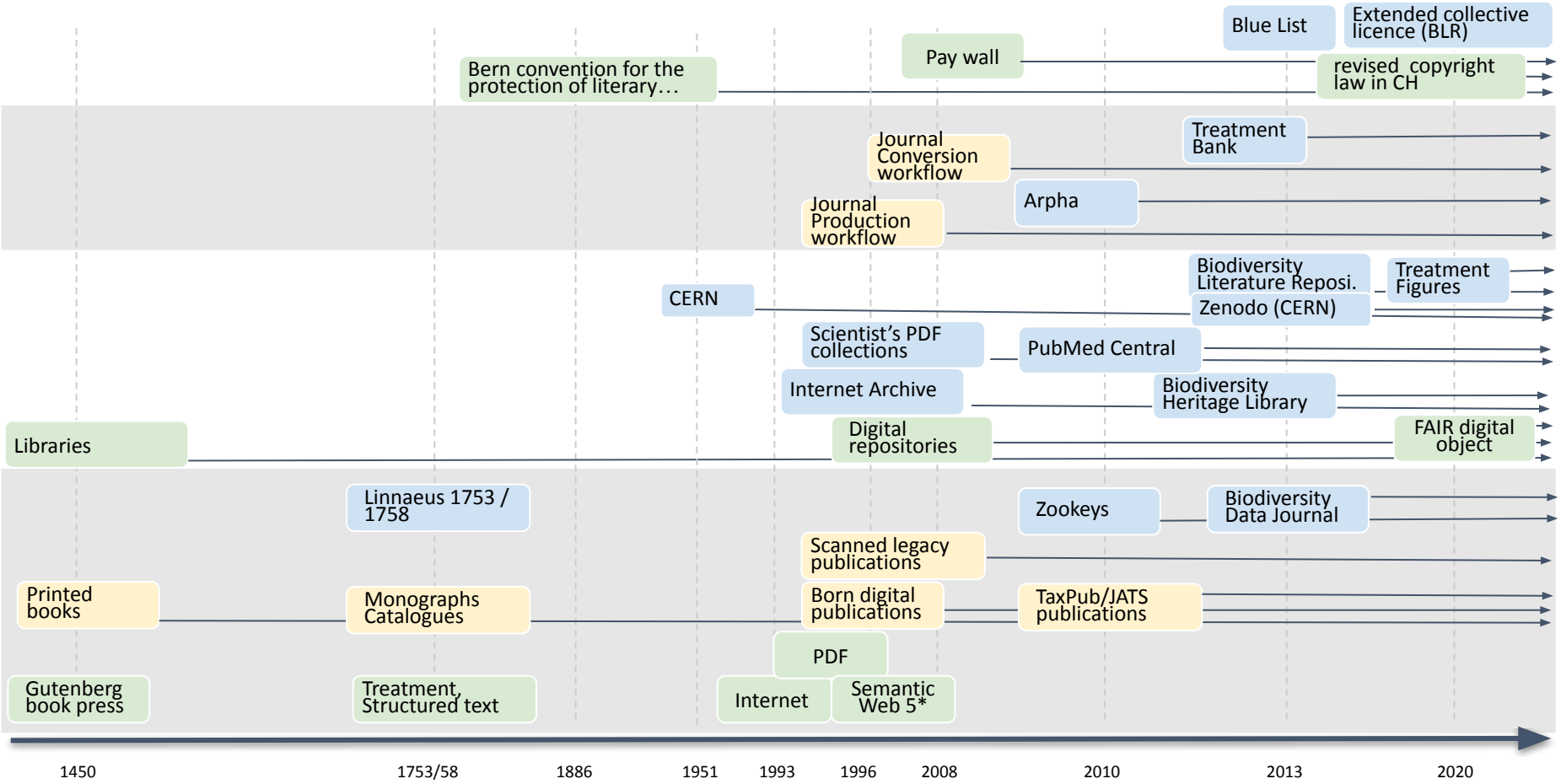
Goal
Communicate the results of charting, cataloguing and describing the Earth' biodiversity

Scope
Scholarly, published research results including all taxa known to science

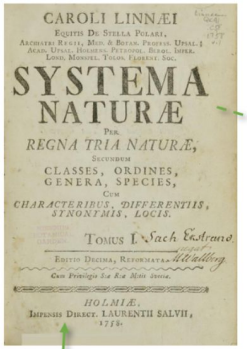
Typical user questions

- How many species do we know?
- What do we know about them?
- Is a taxonomic name available?
- What is the history (synonymy) of a taxonomic name?
- What are the biotic interactions between species?
- How can a taxon be identified?
- How are species related to each other?
- What do I know about a gene?

Brief history of publishing in taxonomy as seen by Plazi



Digital Accessible Knowledge: Treatment


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

SYSTEMA NATURÆ

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

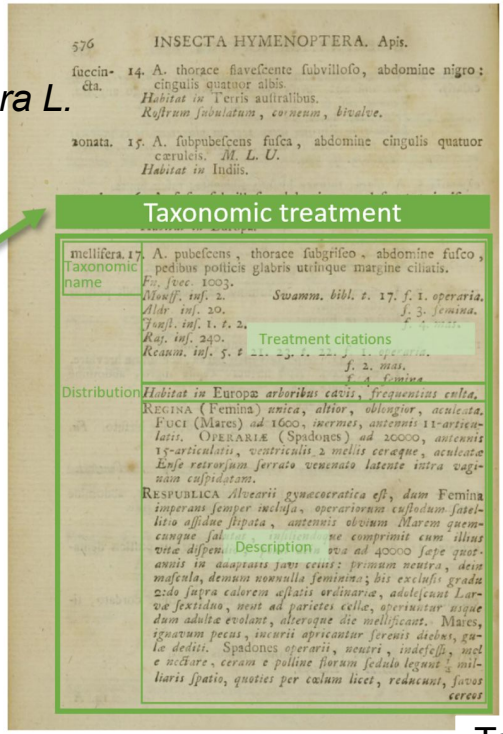
TOMUS I. Sach Castrans

EDITIO DECIMA, REFORMATA.

HOLMIÆ,
IMPENSIS DIRECT. LAURENTII SALVII,
1758.

Linnaeus, 1758
4,819 taxa

Apis mellifera L.


576 INSECTA HYMENOPTERA. Apis.

succinéta. 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

Taxonomic name

mellifera. 17. 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. 2.
Raj. inf. 240.
Reaum. inf. 5. *f.*
f. 2. *mas.*
f. 4. *femina.*

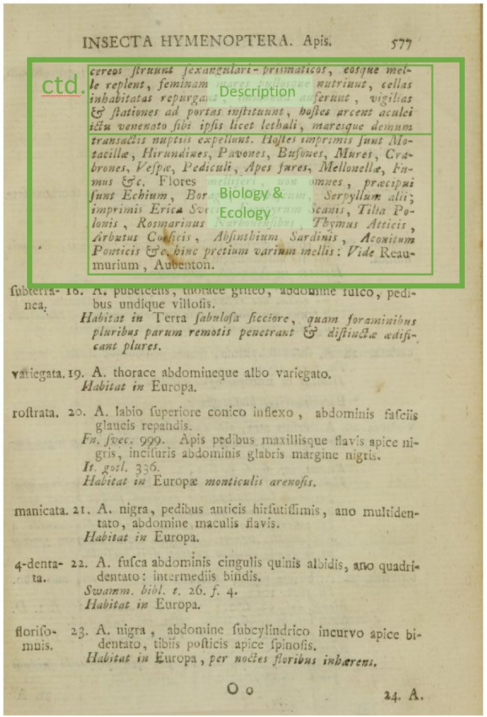
Treatment citations

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

Description

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

REPUBLICA *Alvearii gynæcocratica est, dum* Femina *imperans semper inclusa, operariorum custodum satellitio assidue stipata, antennis obvium Marem quemcunque sal... ne comprimit cum illius vitæ dispen... ova ad* 40000 *sæpe quotannis in duapartita jam cetis: primum neutra, dein mascula, demum nonnulla feminina; his exclusis gradu 2:do supra calorem æstatis ordinariæ, adolescunt Larvæ fextidno, 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 cœlum licet, redeunt, favos cereos*


INSECTA HYMENOPTERA. Apis. 577

ctd.

Description

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

Biology & Ecology

subterra- nea. 18. A. pubescens, thorace griseo, abdomine fusco, pedibus undique villosis.
Habitat in Terra *fabulosa ficcore, quam foraminibus pluribus parum remotis penetrant & diffluctæ 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. guil. 336.
Habitat in Europæ *monticulis arenosis.*

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

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

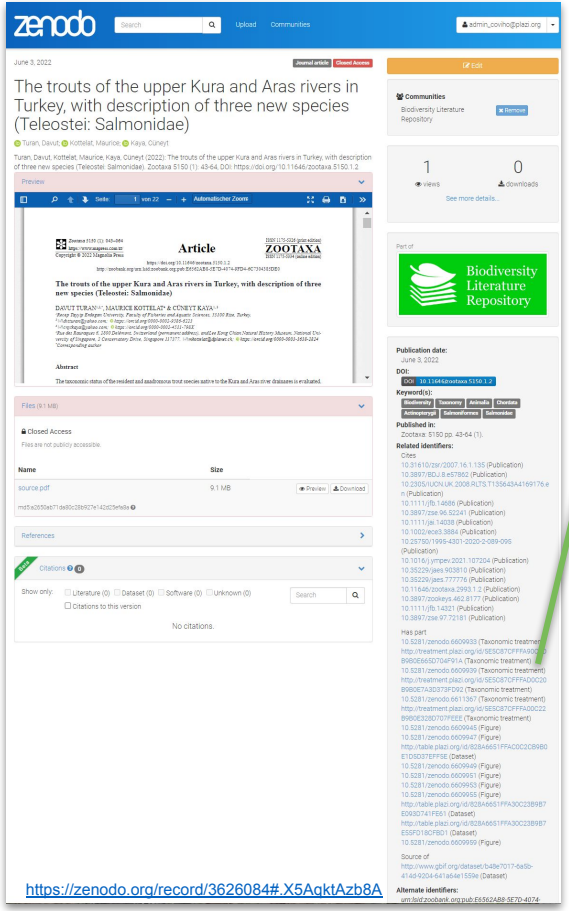
floriso- mnis. 23. A. nigra, abdomine subcylindrico incurvo apice bidentato, tibiis posticis apice spinosis.
Habitat in Europa, *per noctes floribus inhærens.*

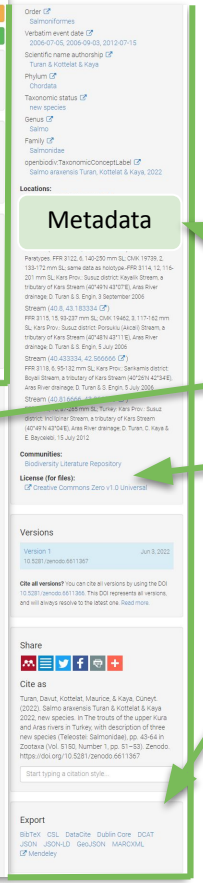
O o 24. A.

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

Treatments as **FAIR digital objects** in the Biodiversity Literature Repository





Findable
Accessible
Interoperable
Reusable

Treatment DOI
Journal article BLR Record
Article DOI

Treatment citation: Catalogue of life

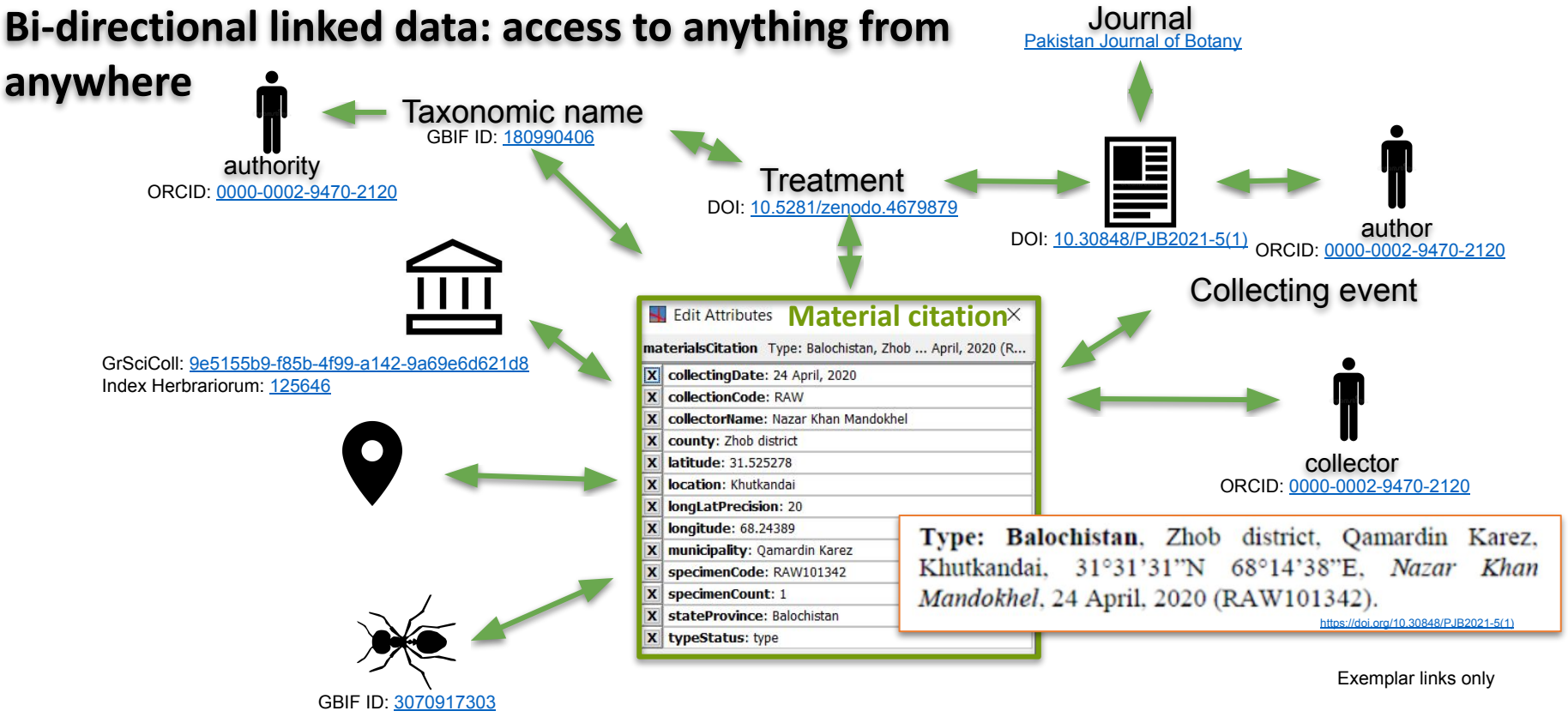
Currently accepted name

Synonymized taxonomic name

Original name or new combination

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?

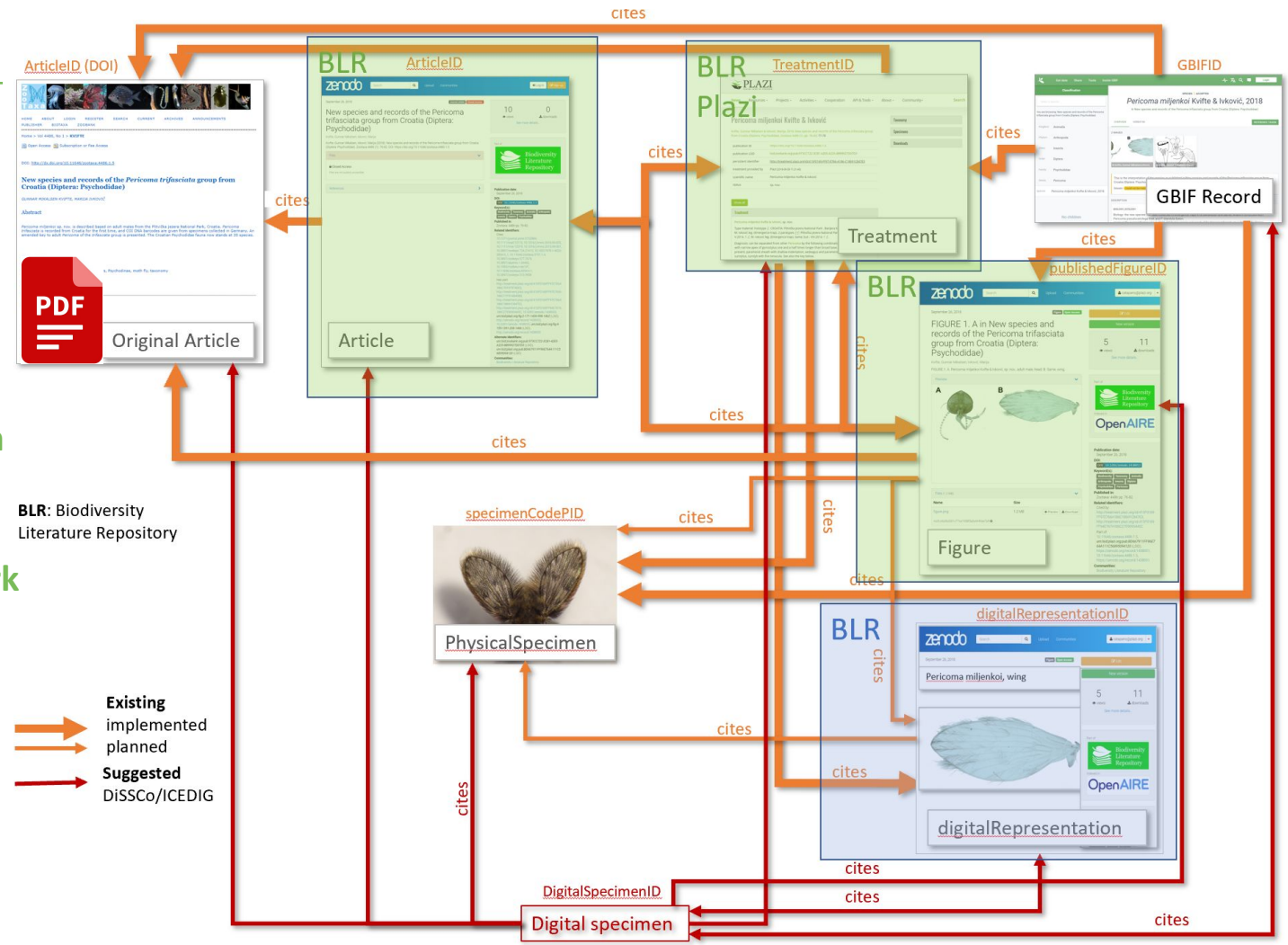
Reuse of taxonomic data: GBIF

A material citation is an identified specimen, linked to the respective treatment

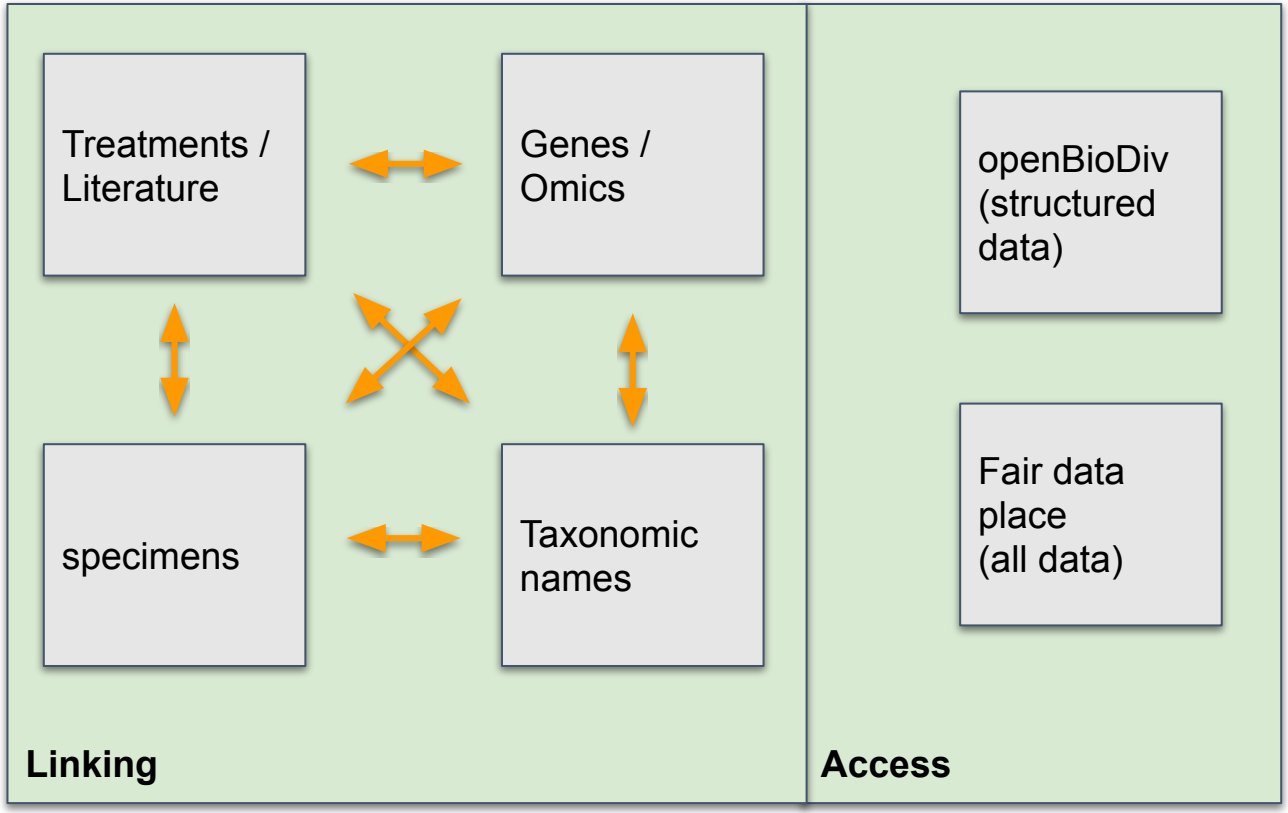
GBIF species: https://www.gbif.org/species/1703942
GBIF occurrence: https://www.gbif.org/occurrence/3015409563
BLR Taxonomic treatment: https://doi.org/10.5281/zenodo.4397319
BLR Article: https://zenodo.org/record/4397315
PDF: https://doi.org/10.11646/zootaxa.4820.1.4

Focus on data in publications: What's inside?

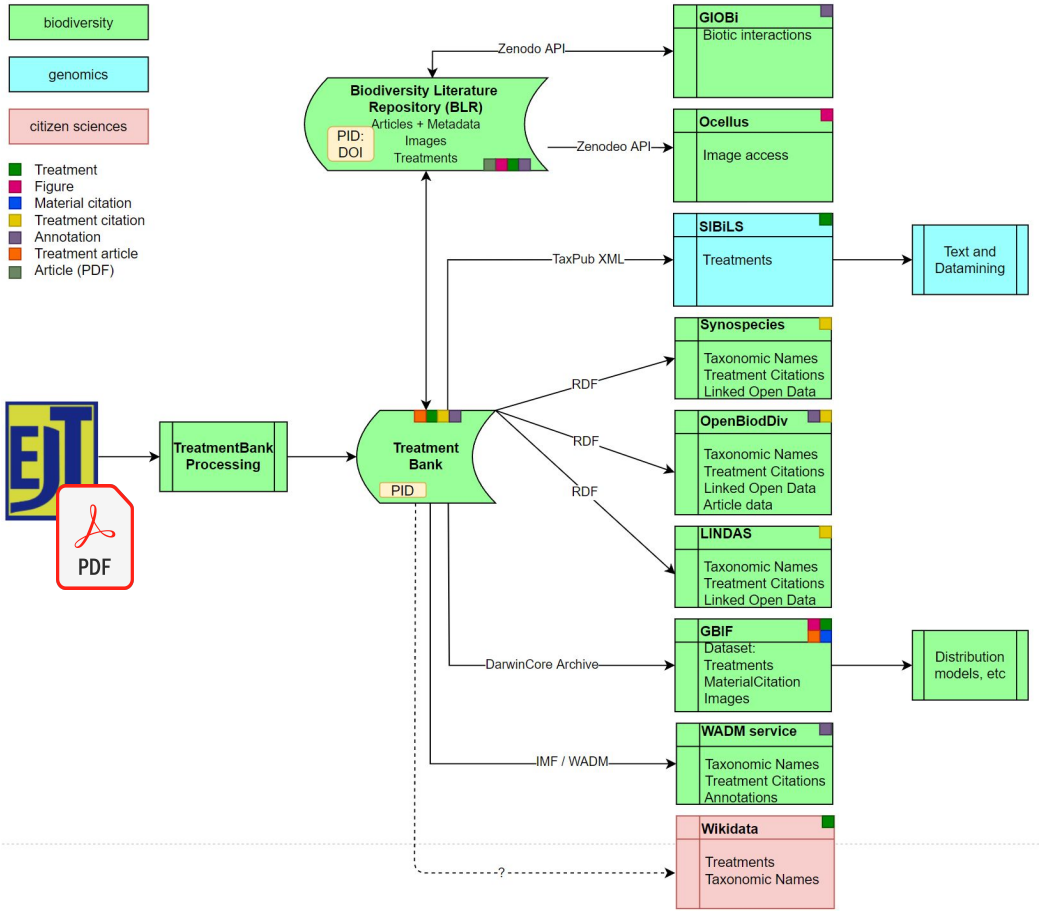
Citation network



BiCIKL: Linking and access



Re-use of scholarly publication's data



Infrastructures maintained by Plazi

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

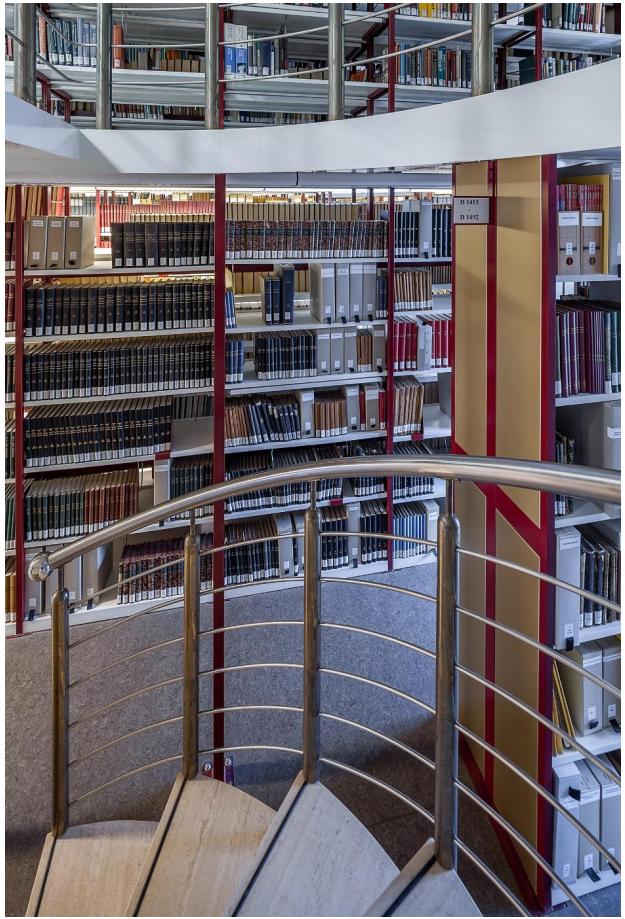
Biodiversity Literature Repository

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

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

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)

Plazi: Goal

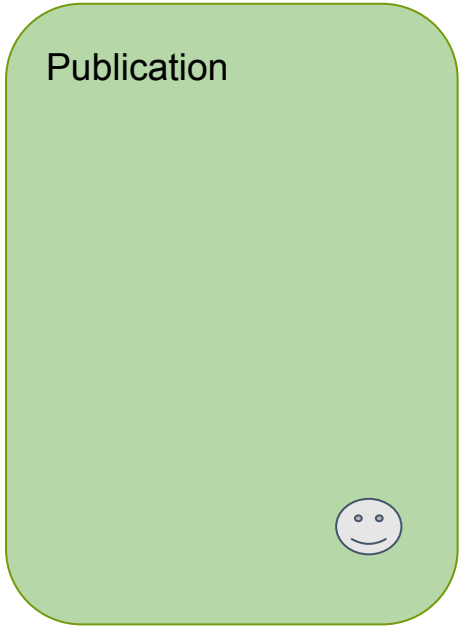
Discovering known biodiversity

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)

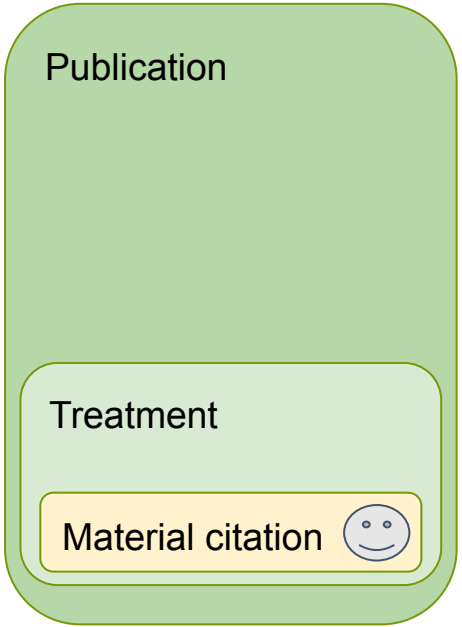
- Data understandable by human and actionable by machine reflecting the growth of our knowledge and interdependence.
- Editorial structure including textflow, paragraphs, sections such as title, authors and affiliations, materials and methods, etc., tables, figures, bibliographic references and their 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



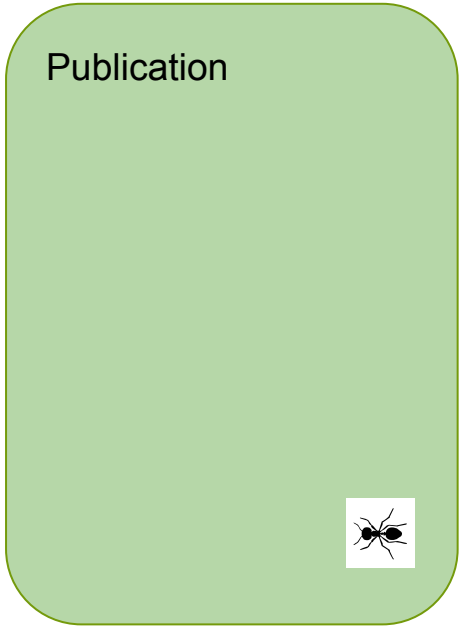
Person: **role**
Public. author, authority, collector

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

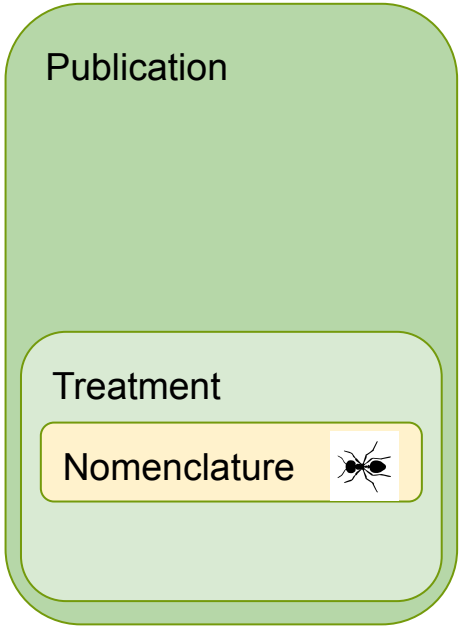


Person: **role**
Publication
Treatment
Material citation collector

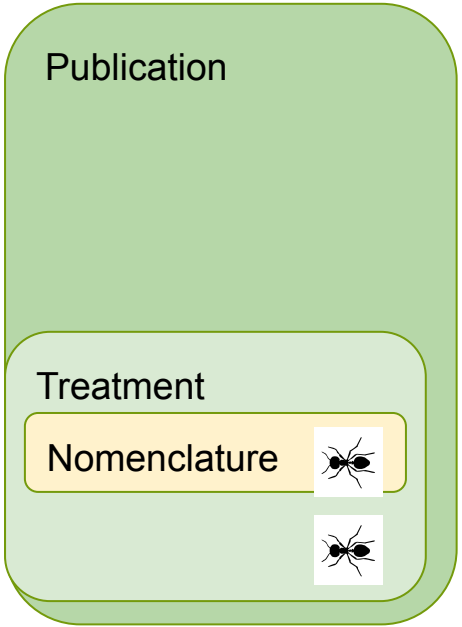
Annotations are context defining the role of named entities



Taxonomic name
Publication is part of



Taxonomic name
Publication is part of
Treatment (nomenclature) is about



Taxonomic name
Publication is part of
Treatment (nomenclature) is about
Treatment body is related to

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

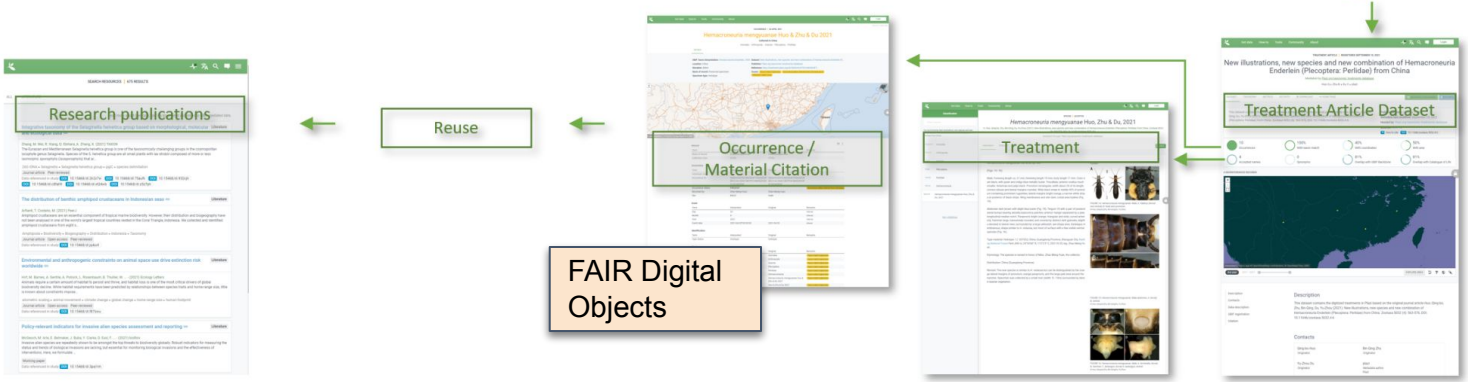
Access / Metrics
- User Interface
- API
- Dashboards

SynoSpecies
Taxonomic names
Treatments

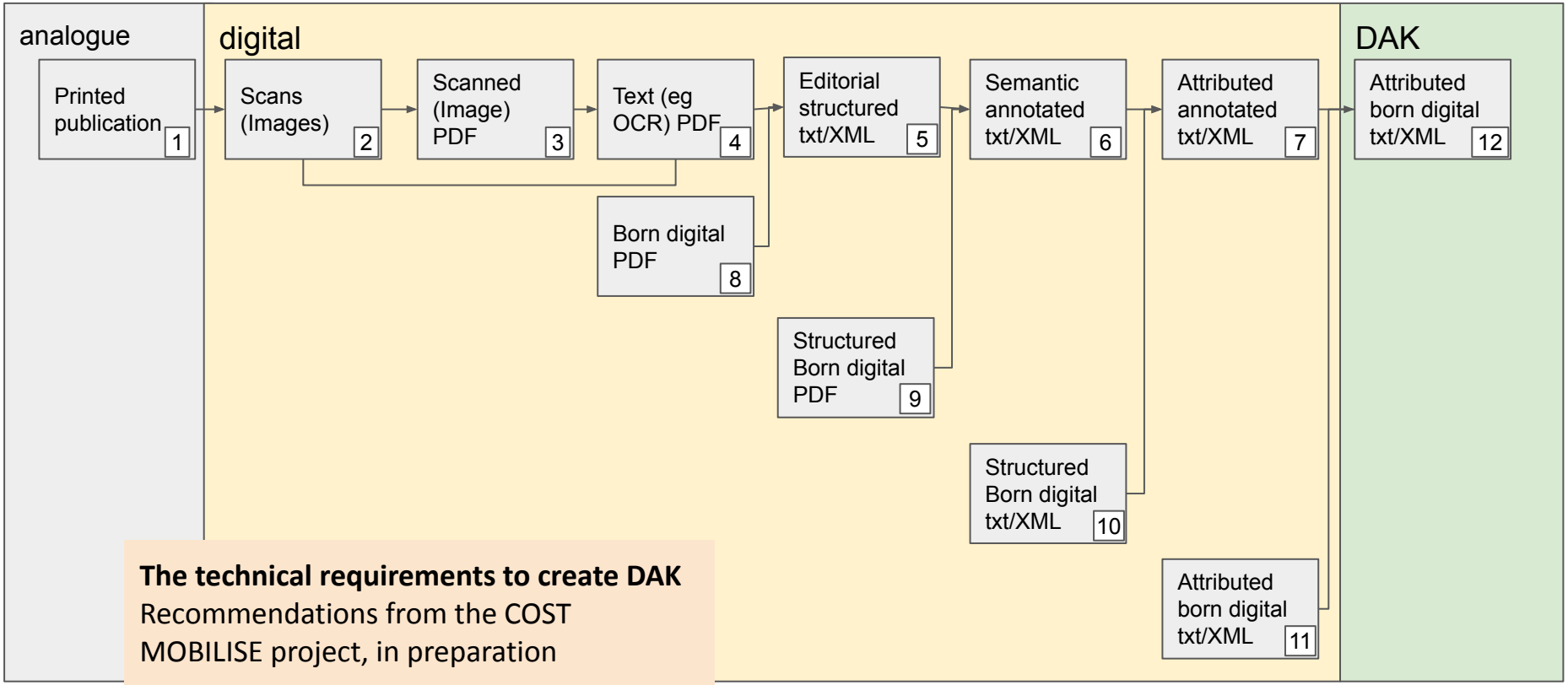
Disseminate
- GBIF
- NCBI

Reuse

GBIF



Origin of Digital Accessible Knowledge (DAK)



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

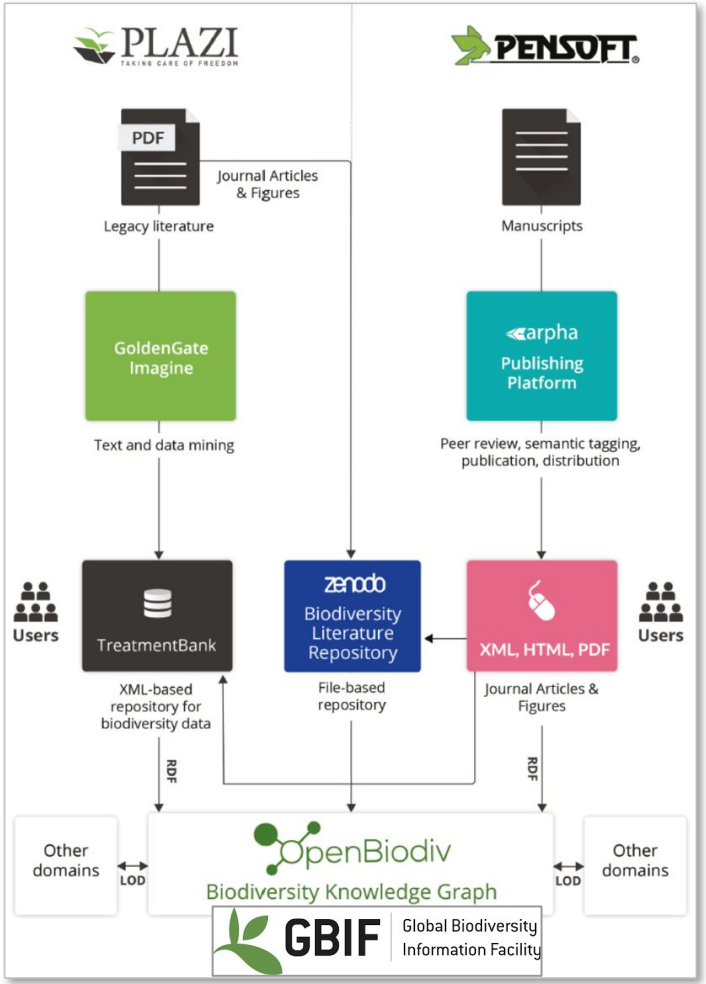
Data conversion

Challenges:

- Publication representation (e.g. Journal vs issues vs article)
- Metadata
- Page image scan quality
- OCR artefacts from text flow, incomplete text recognition to misspelled words
- PDF formats
- Treatments
- Taxonomic names
- Material citations
- Treatment citations
- Treatment internal structure
- Making implicit links explicit

The future will be open, semantically enhanced publishing

Legacy publications



Prospective publishing

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

Thank you!

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

Introduction to digitizing taxonomic literature with Plazi DOI

Donat Agosti

agosti@plazi.org

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).