Doi: 10.5281/zenodo.6777916

Beyond the print and PDF prisons: Data about biodiversity want to be free

June 30, 2022

World Biodiversity Forum

Davos, Switzerland

Donat Agosti

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

Plazi, Switzerland

SIBILS
SWISS INSTITUTE OF BIOINFORMATICS
LITERATURE SERVICES

swissuniversities

101007492

ARCADIA

BiCIKL

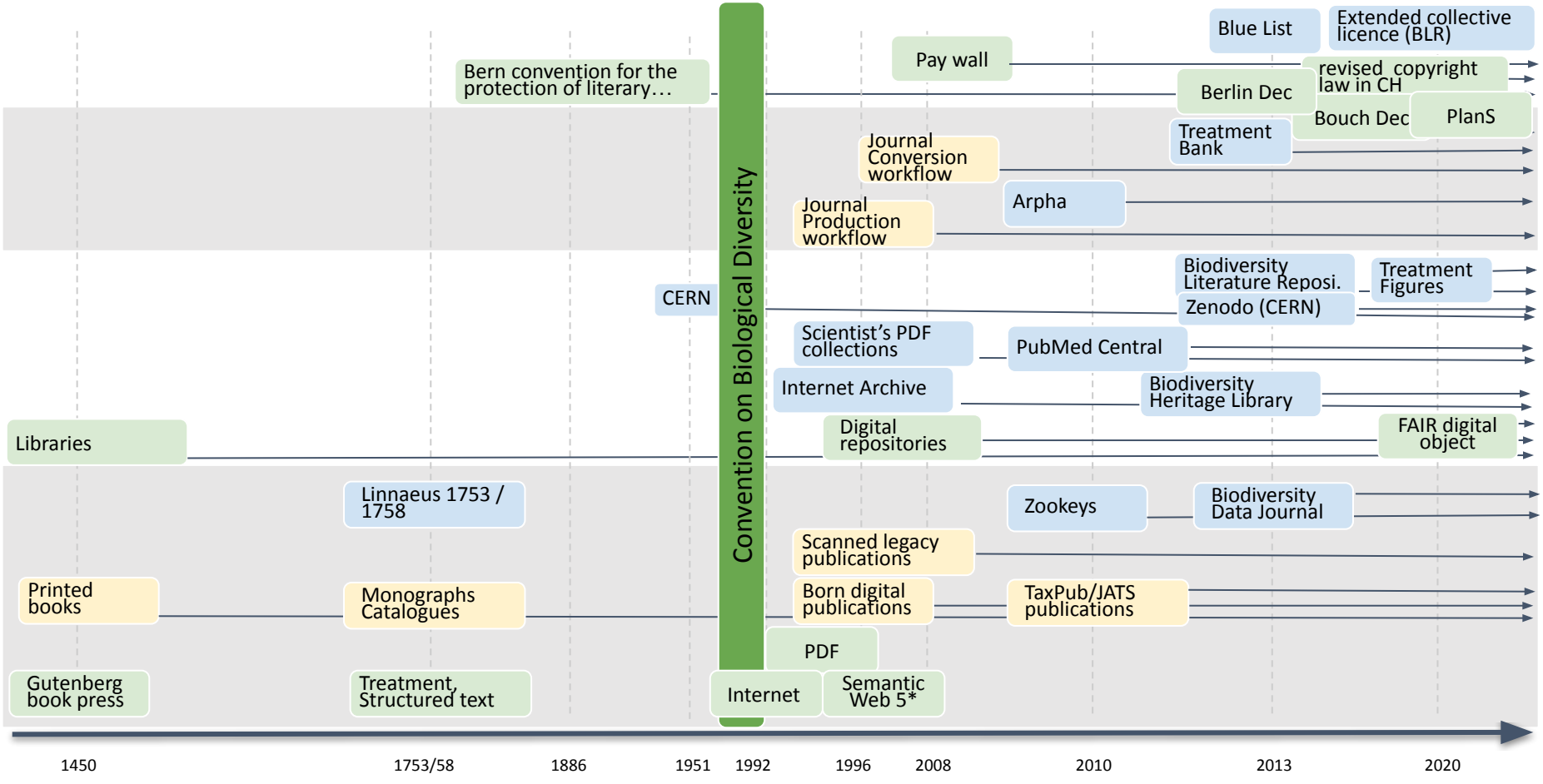
eBioDiv

Biodiversity Literature Repository

mobilise Mobilise Action
Mobilising Data, Policies and Experts in Scientific Collections

TreatmentBank

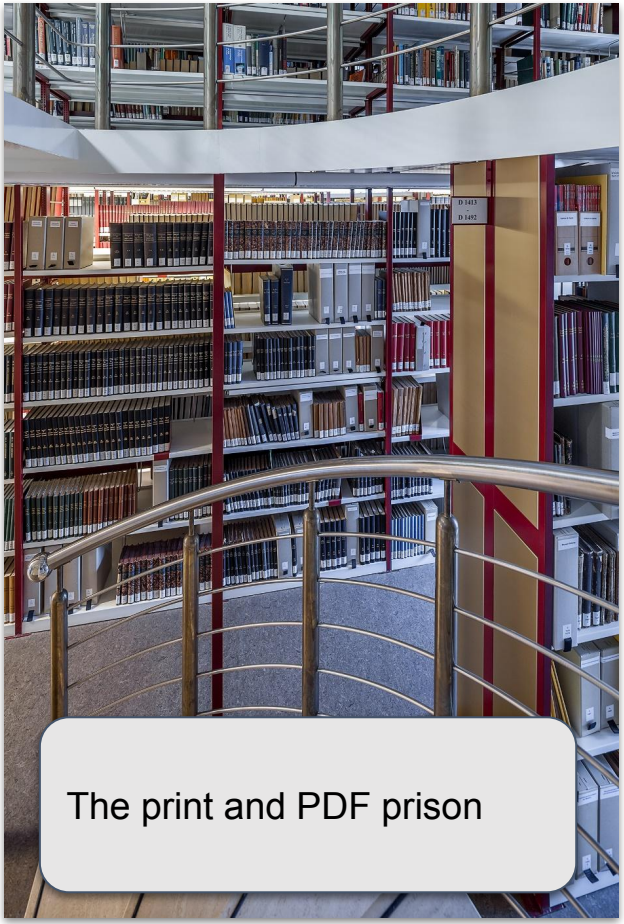
Brief history of publishing in biodiversity / taxonomy



United Nations

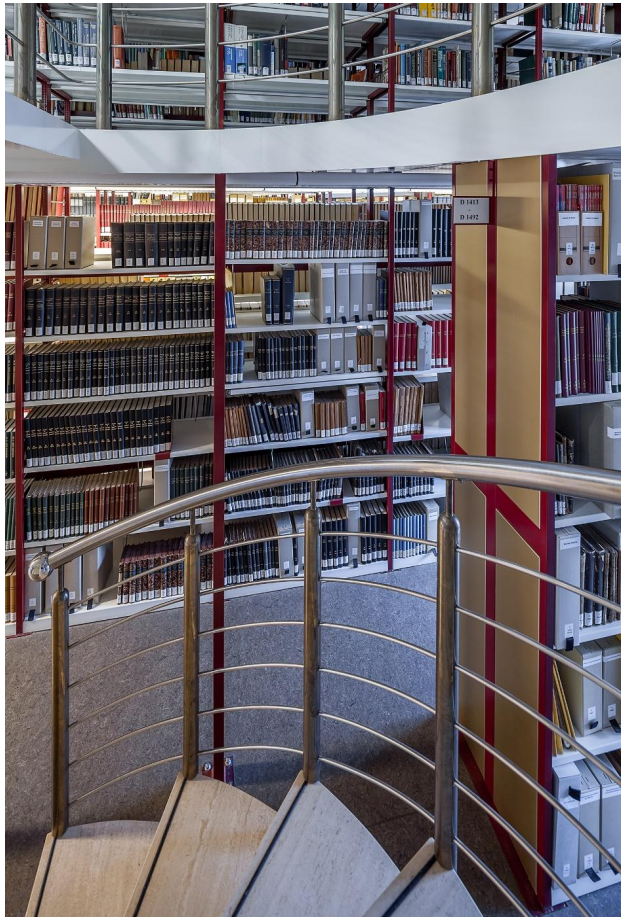
Article 7. Identification and Monitoring
How many species do we lose?
How many species do we know?
How many species are on Earth?

Article 6. General Measures for Conservation and Sustainable Use
What do we know about the species?



BGBM.org

The challenge



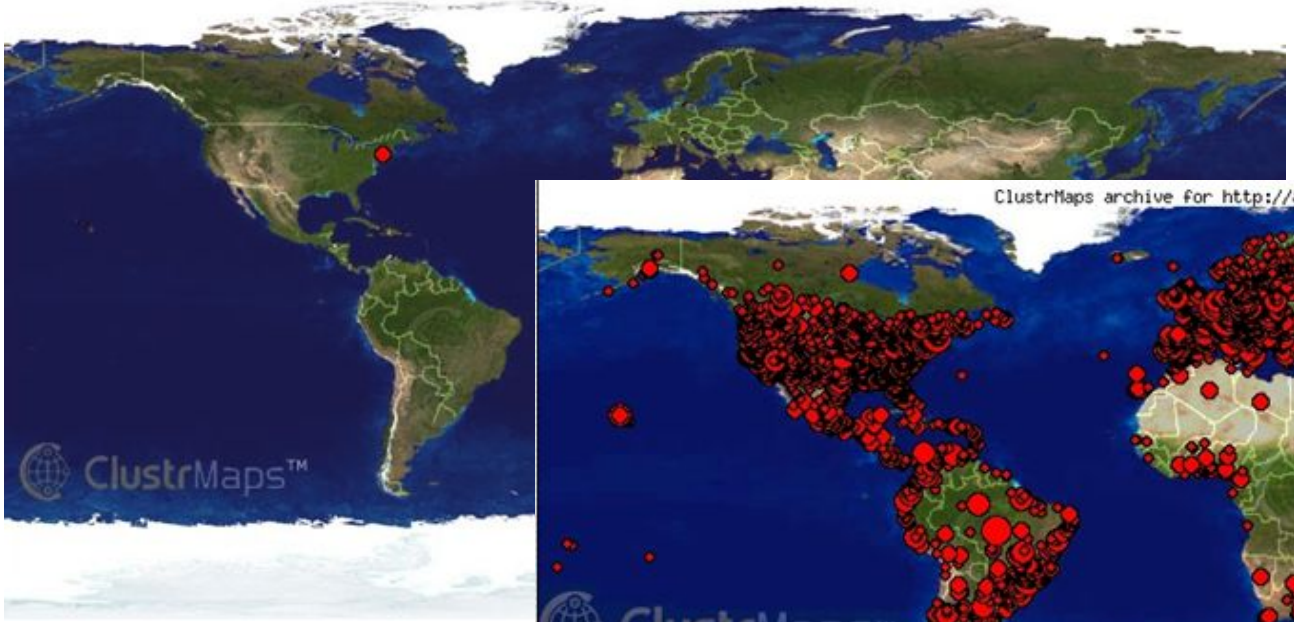
Known biodiversity knowledge

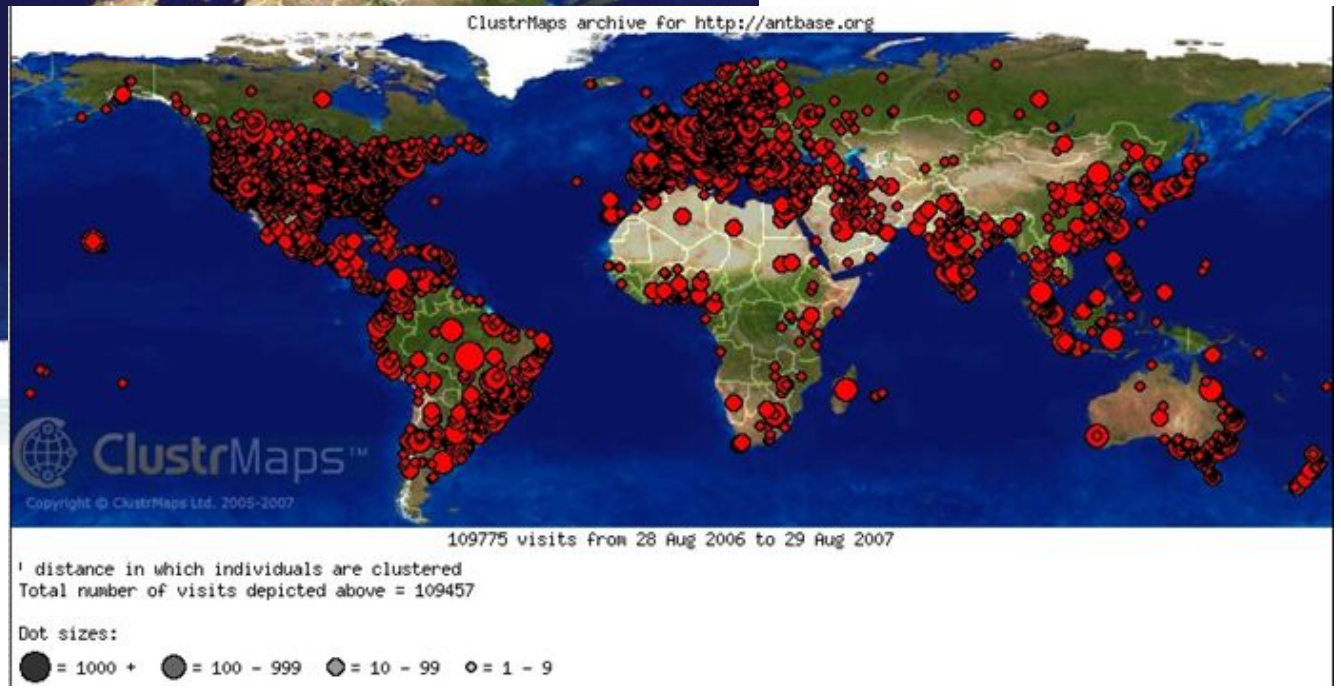
- Empiric science: all results published
- 500,000,000+ printed pages
- >> 1,000 journals publishing taxonomic content
- > 1,900,000 species described
- > 20,000,000+ taxonomic treatments
- approx. 18,000 new species discovered / year
- >> Millions of specimens identified by specialists (material citations)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

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

The impact of WWW in the 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).

Plazi: Goal

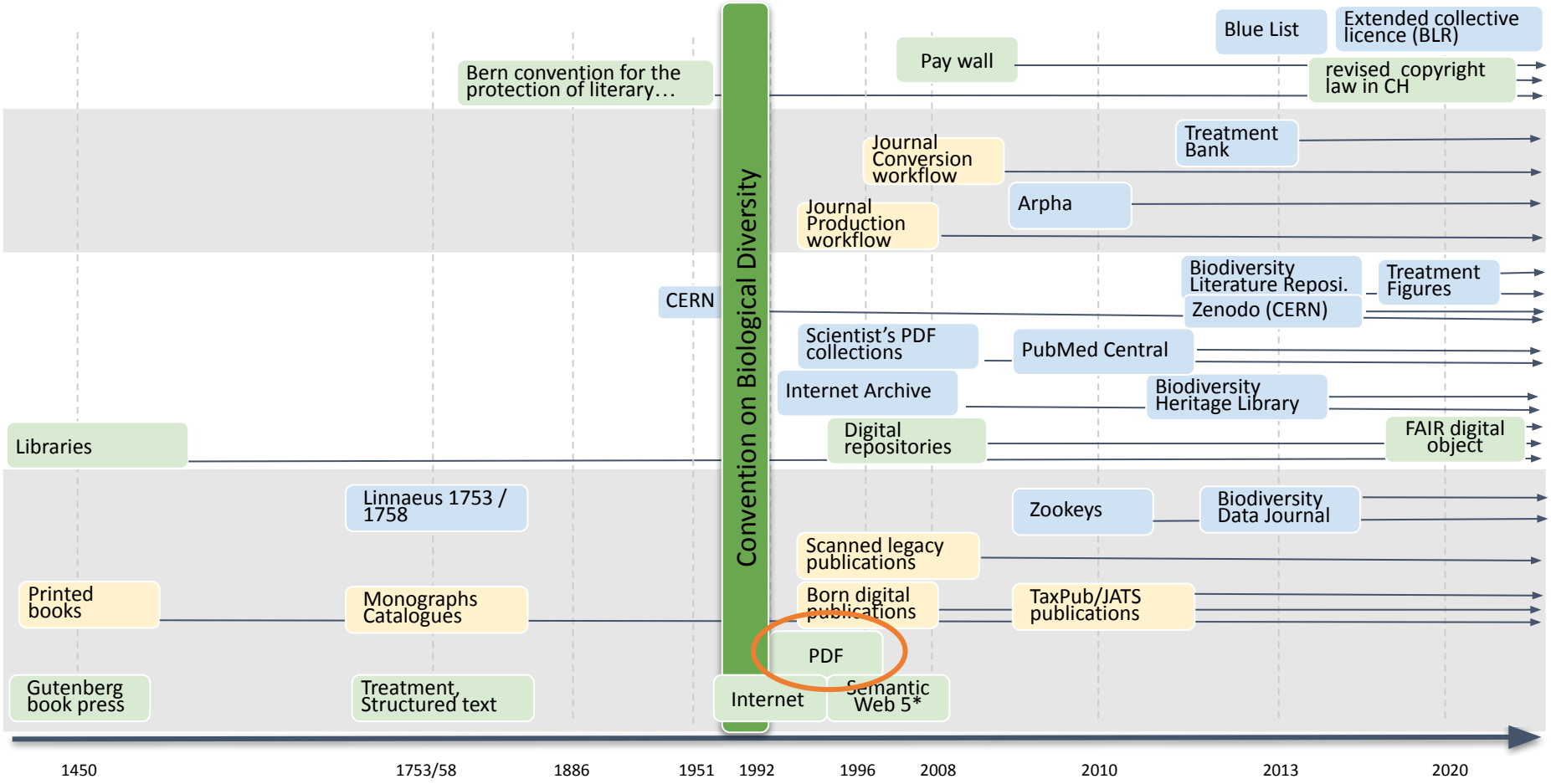
Discovering known biodiversity

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.

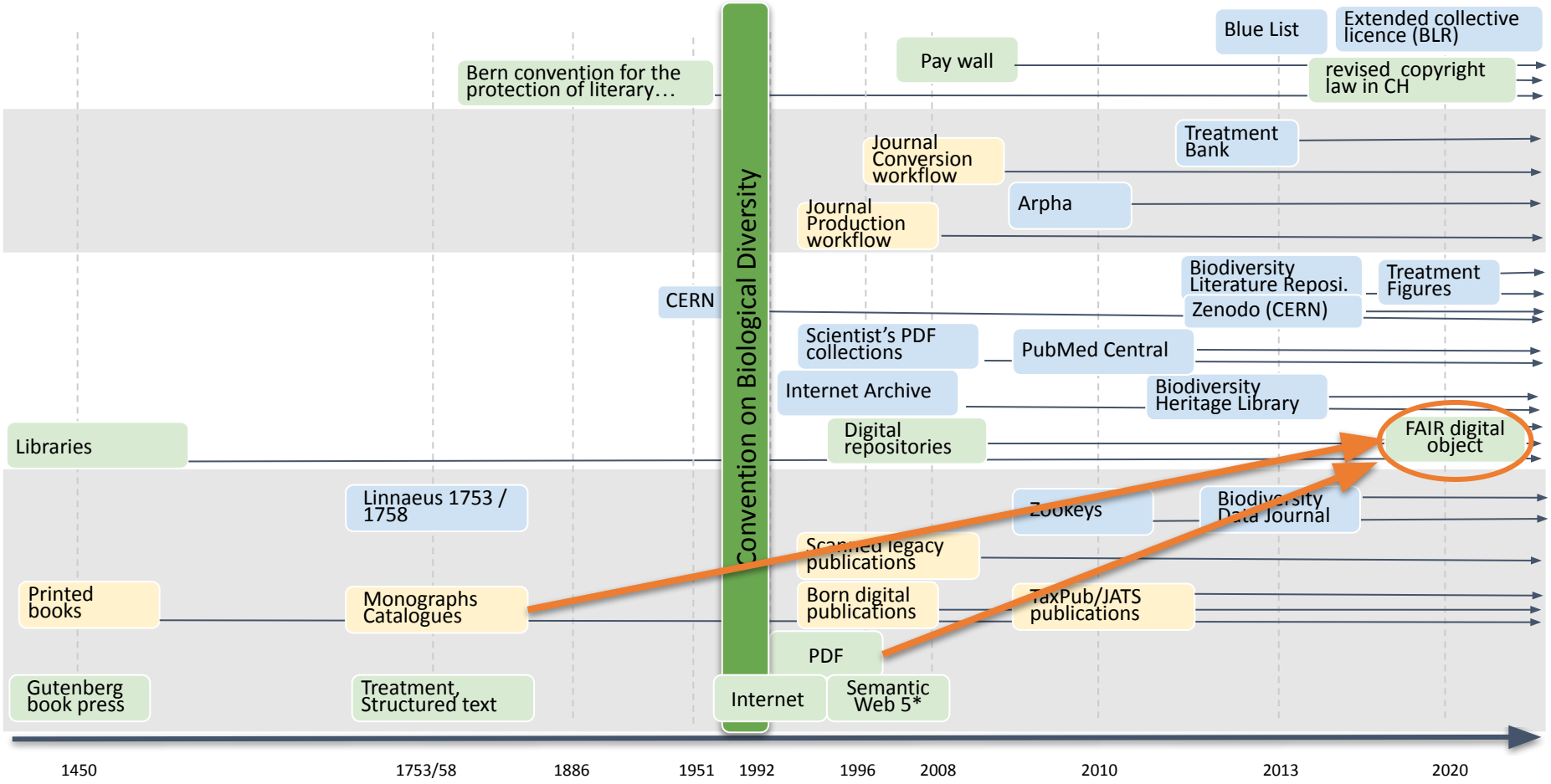
Digital accessible knowledge in biodiversity (DAK)

- Data understandable by human and actionable by machine reflecting the growth of our knowledge and interdependence of biodiversity.
- 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

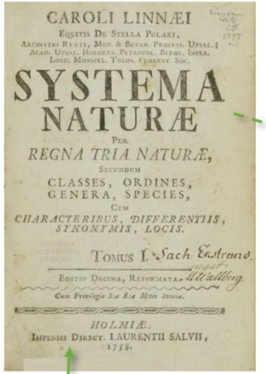
Brief history of publishing in taxonomy as seen by Plazi



Brief history of publishing in taxonomy as seen by Plazi

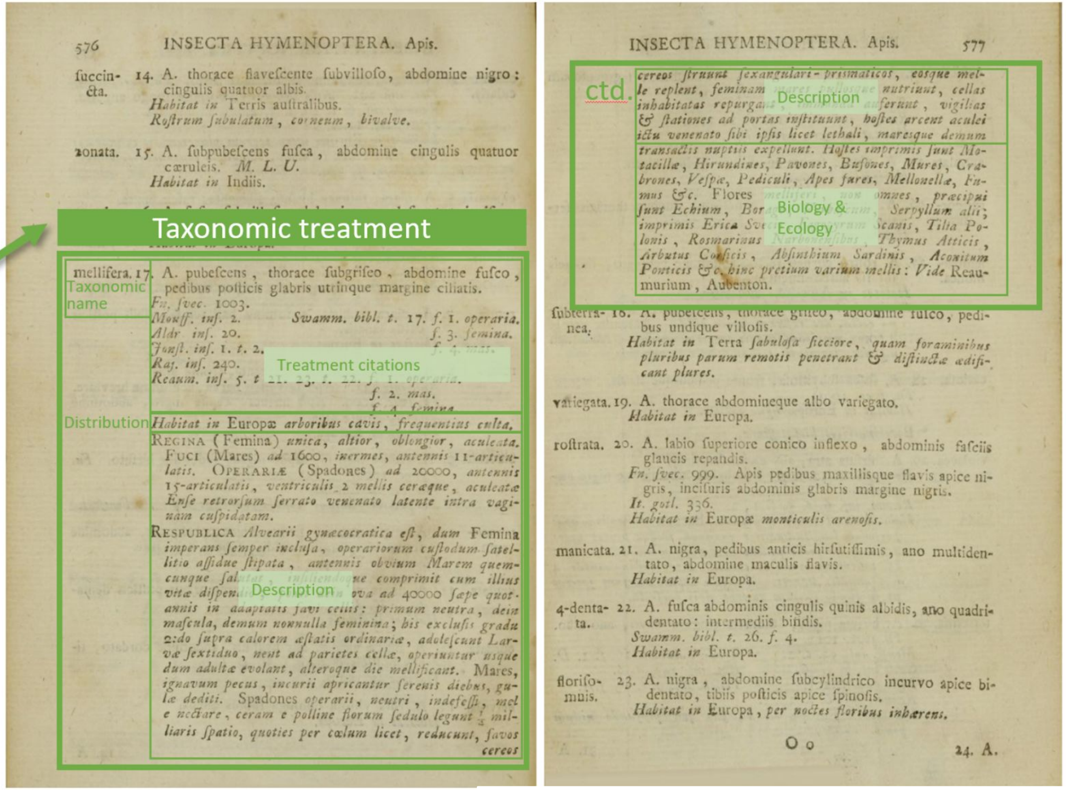


Digital Accessible Knowledge: Treatment



Linnaeus, 1758
4,819 taxa

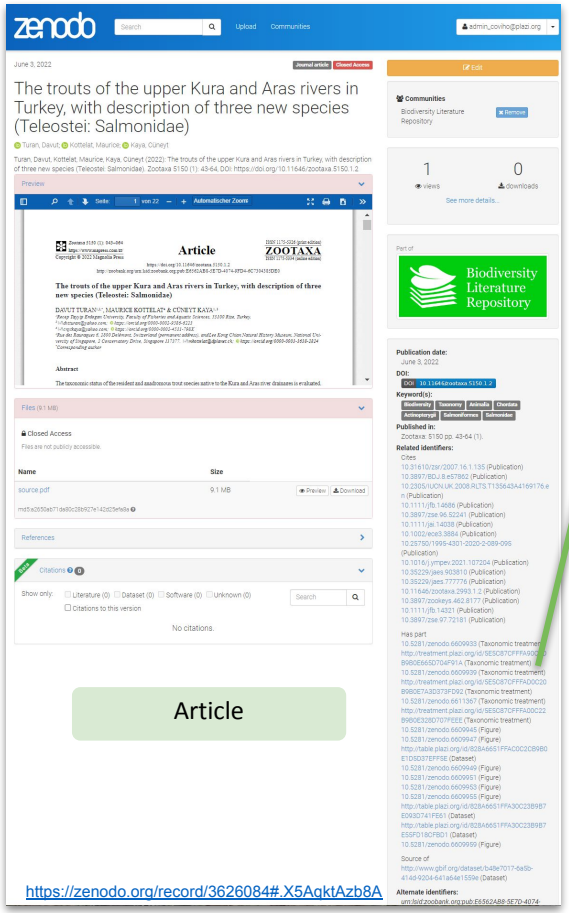
Apis mellifera L.

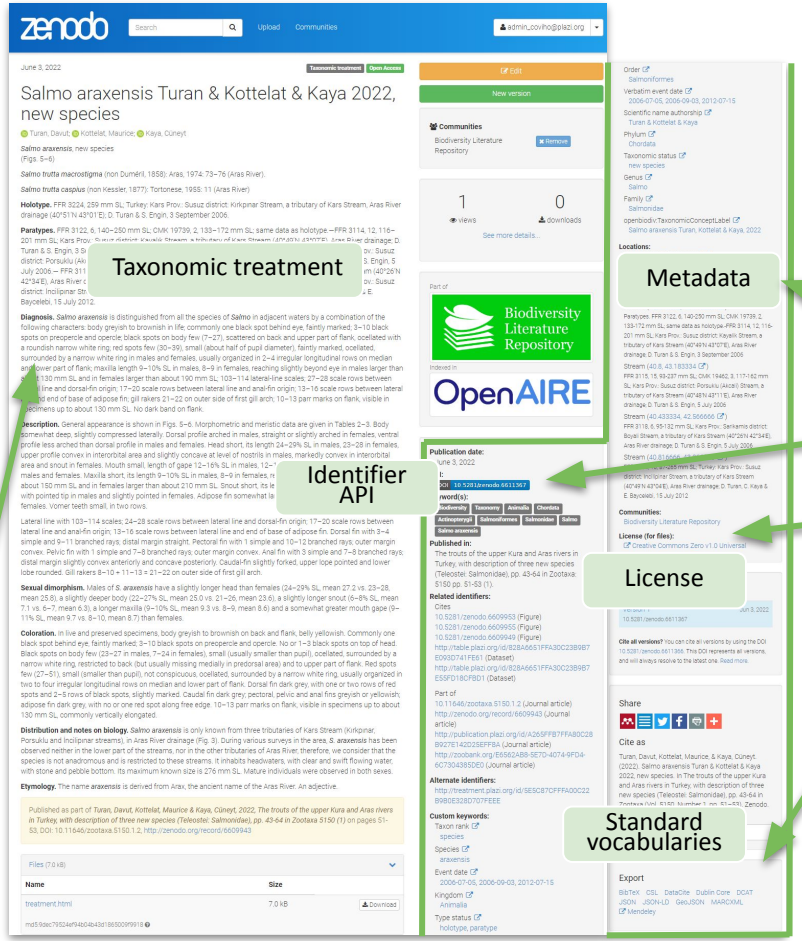


Each type has a published taxonomic treatment
Each taxonomic treatment is multiple times augmented

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

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

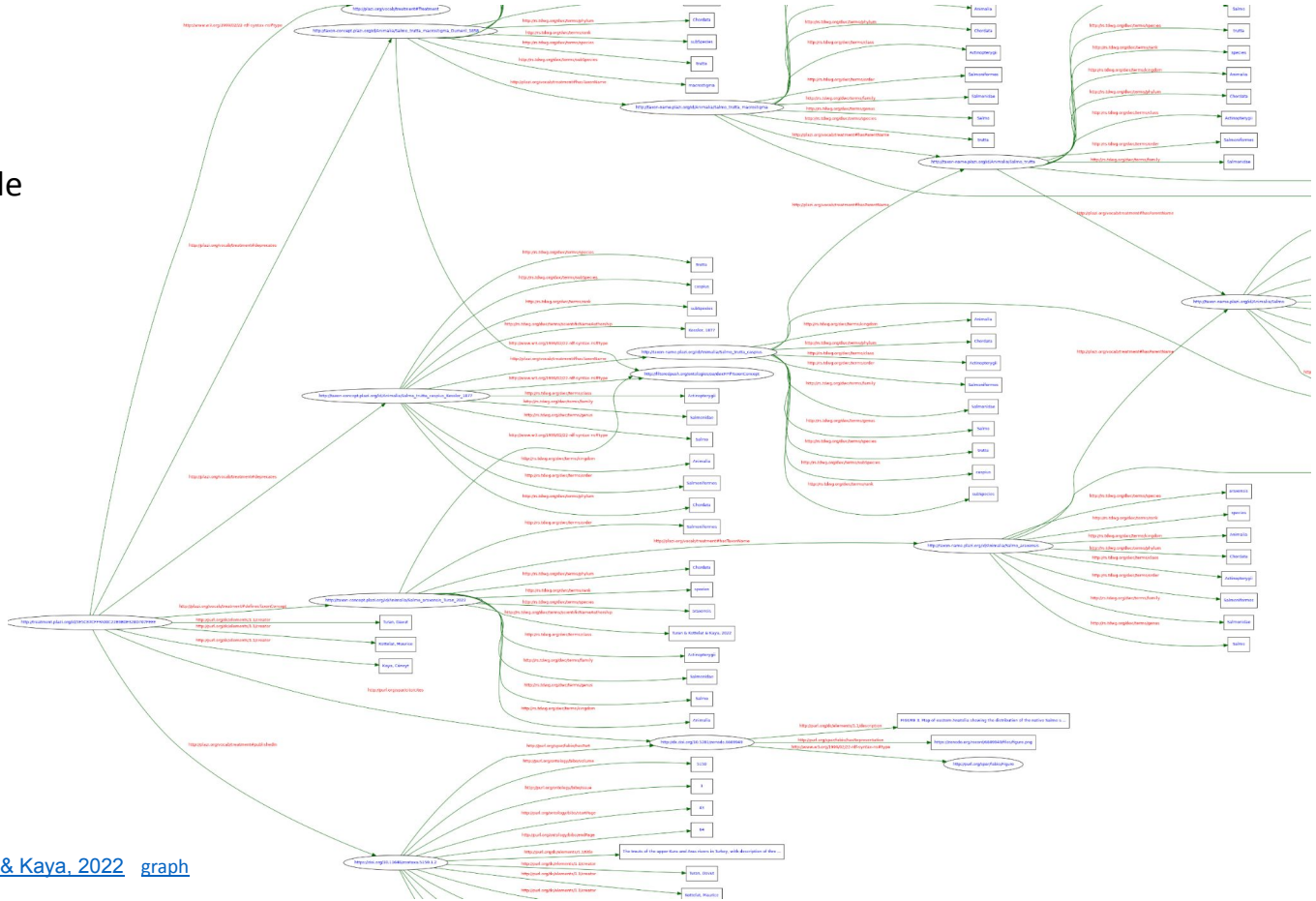




Treatment DOI
Journal article BLR Record
Article DOI

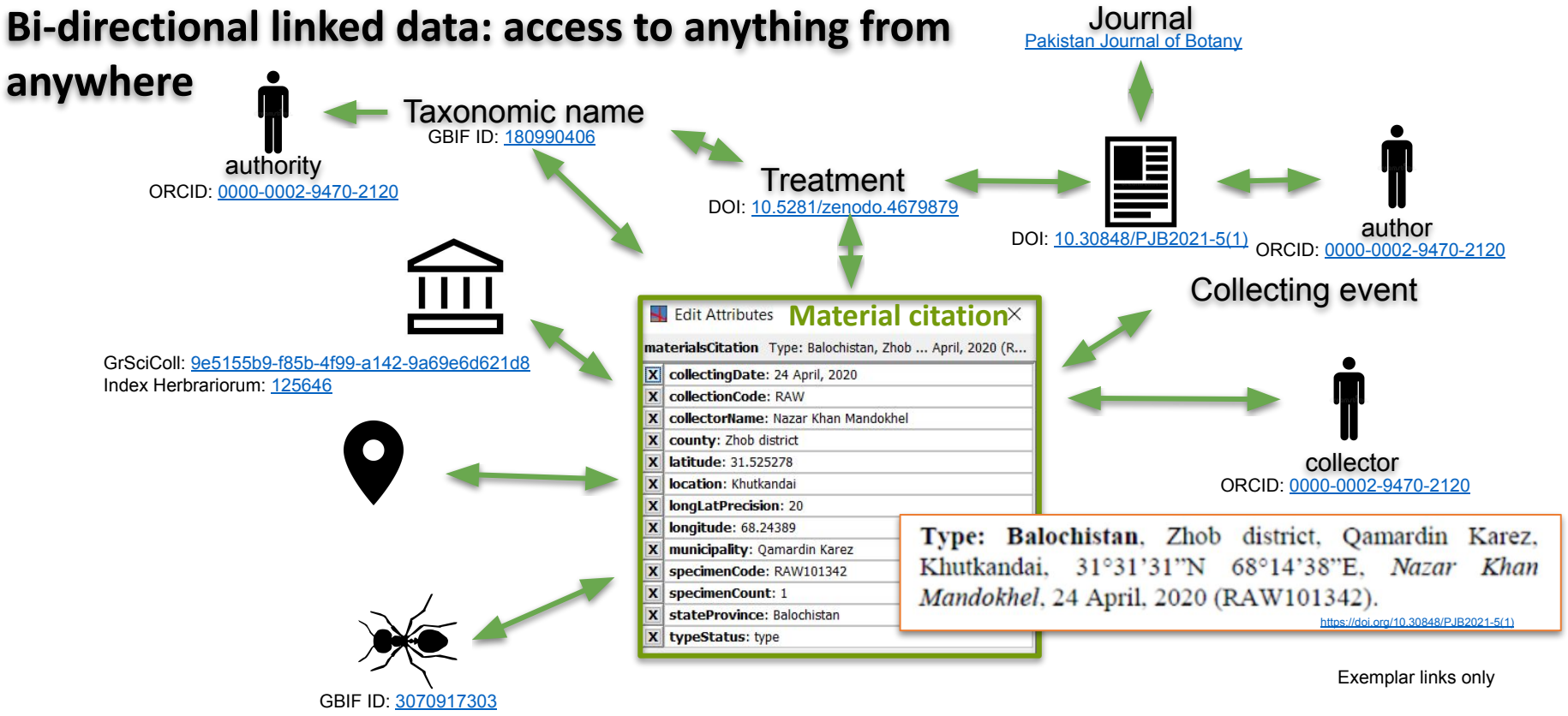
Treatments as **FAIR digital objects** in the Biodiversity Literature Repository

Visualization of the knowledge graph of a single treatment (part)



Example: Salmo araxensis, Turan & Kottelat & Kaya, 2022 graph

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

Treatment citation: Catalogue of life

Currently accepted name

Synonymized taxonomic name

Original name or new combination

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

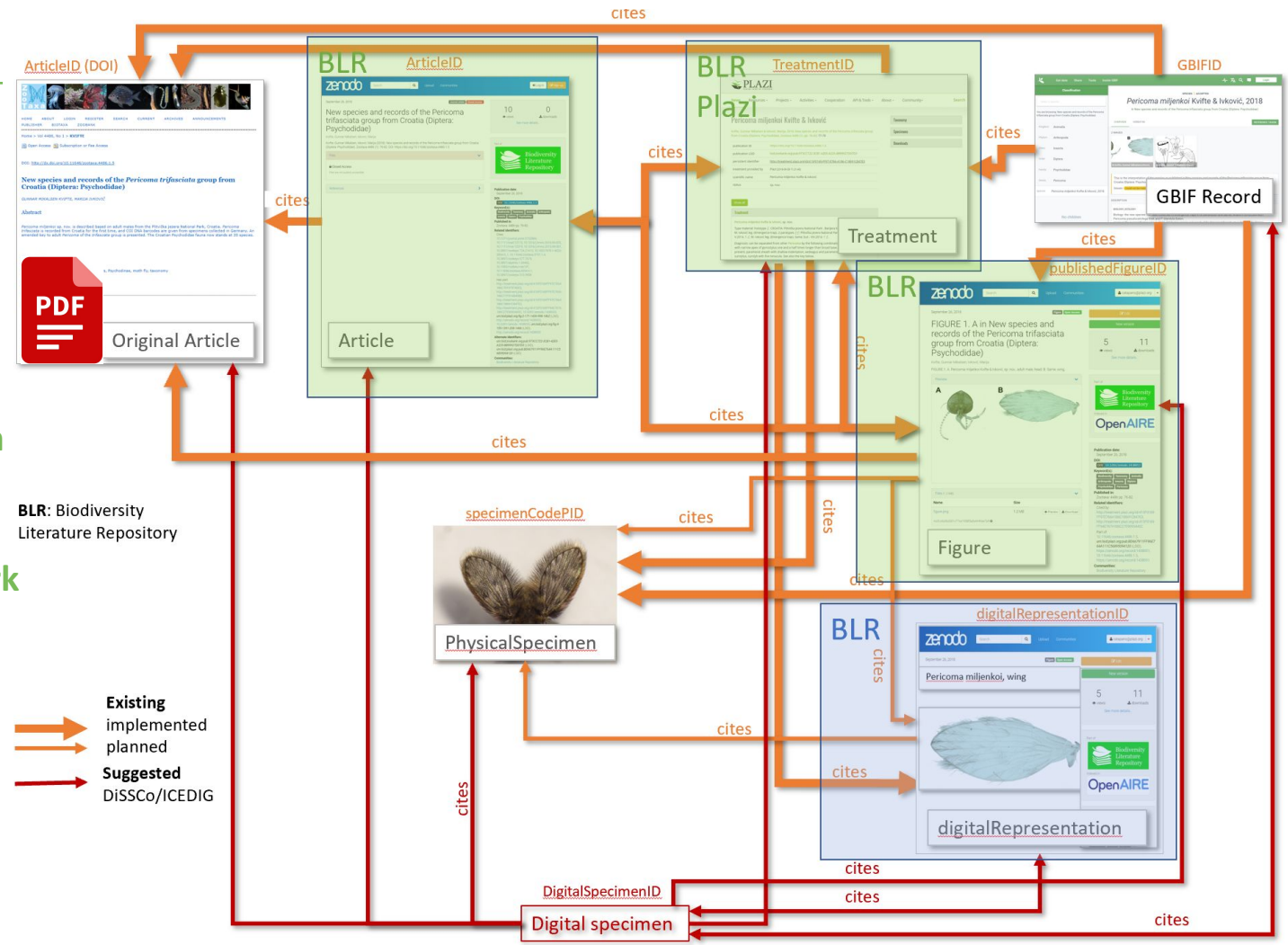
Open access to FAIR data: A visual index to taxonomic publications

Beyond PDF...

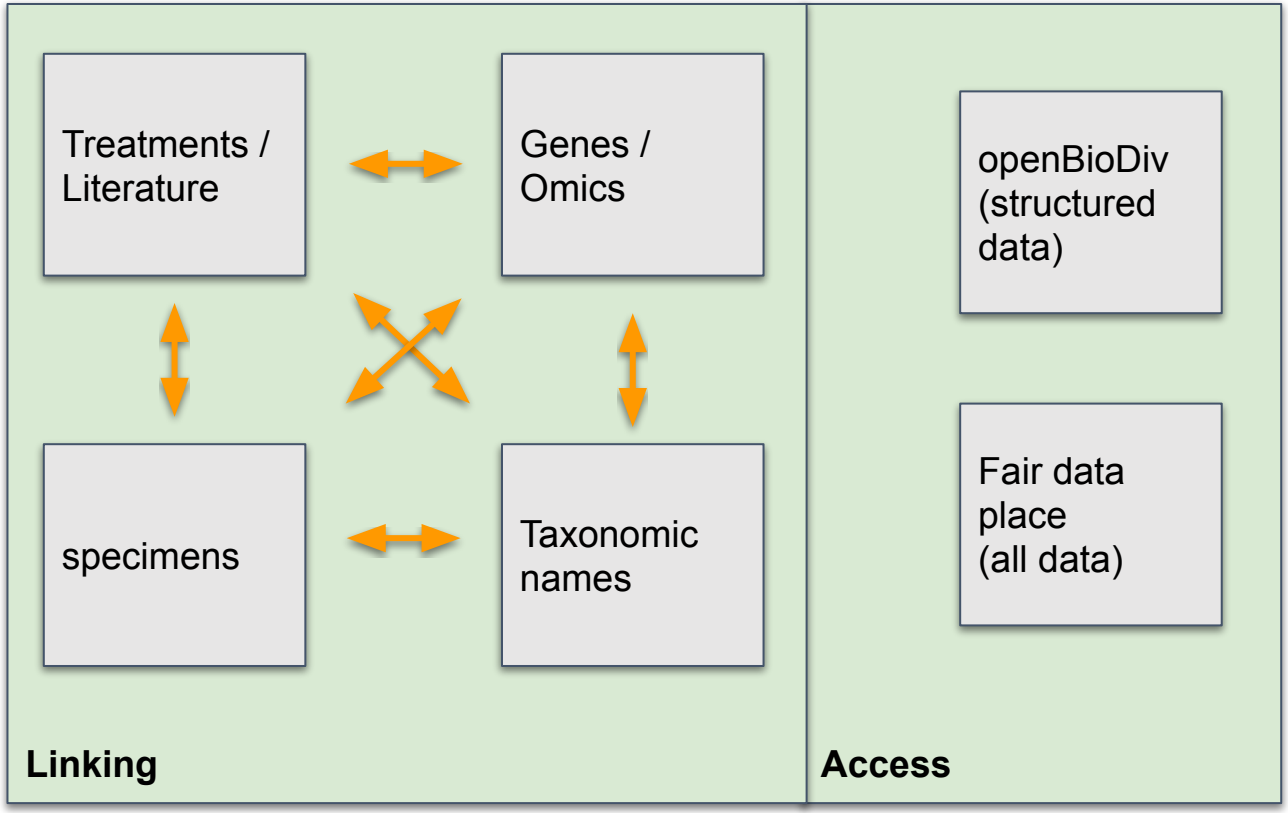
https://ocellus.info/

Focus on data in publications: What's inside?

Citation network



BiCIKL: Linking and access



Plazi workflow

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

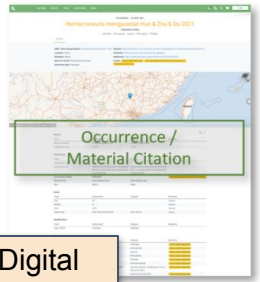
Access / Metrics
- User Interface
- API
- Dashboards

SynoSpecies
Taxonomic names
Treatments

Disseminate
- GBIF
- NCBI

Reuse

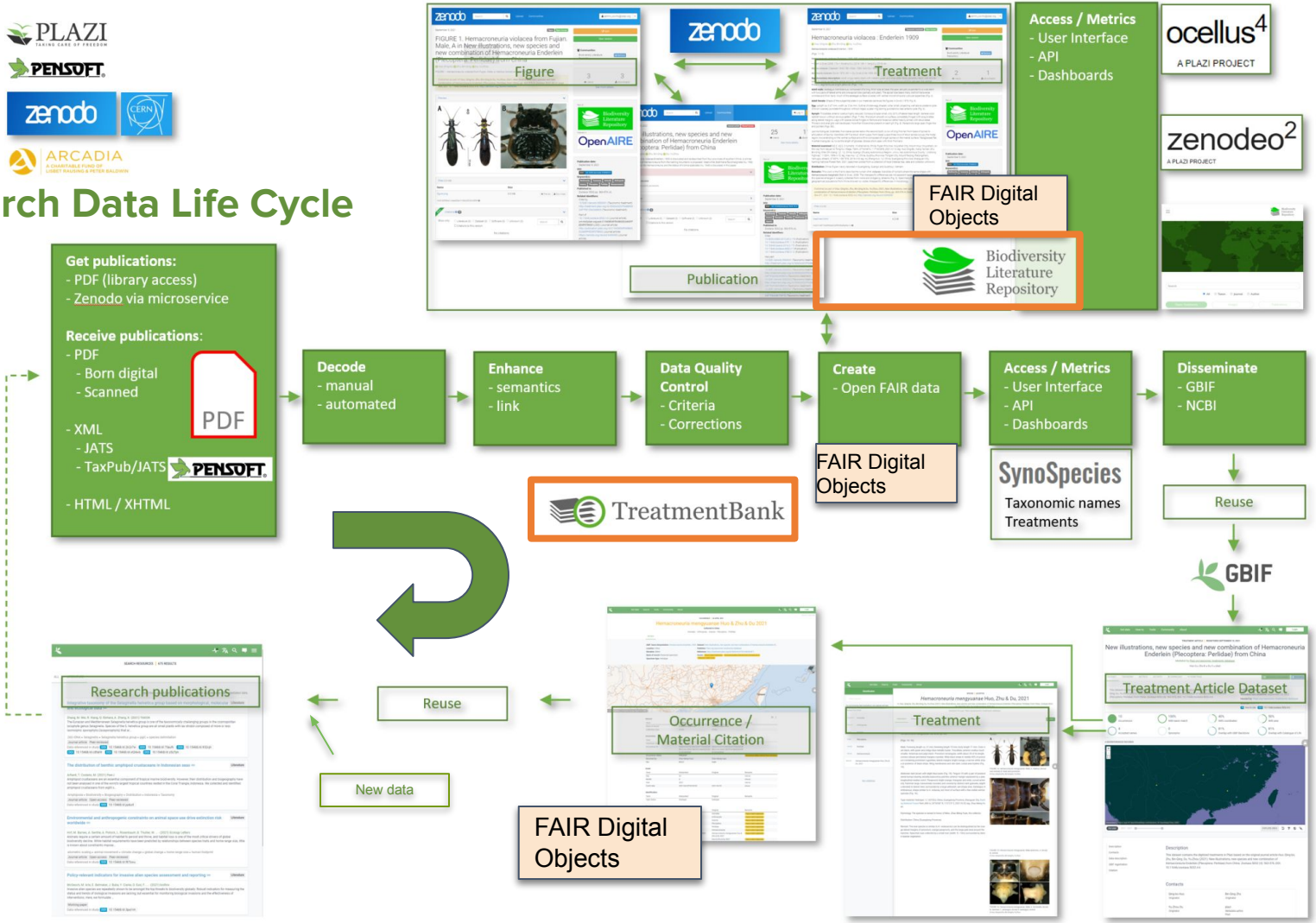
GBIF



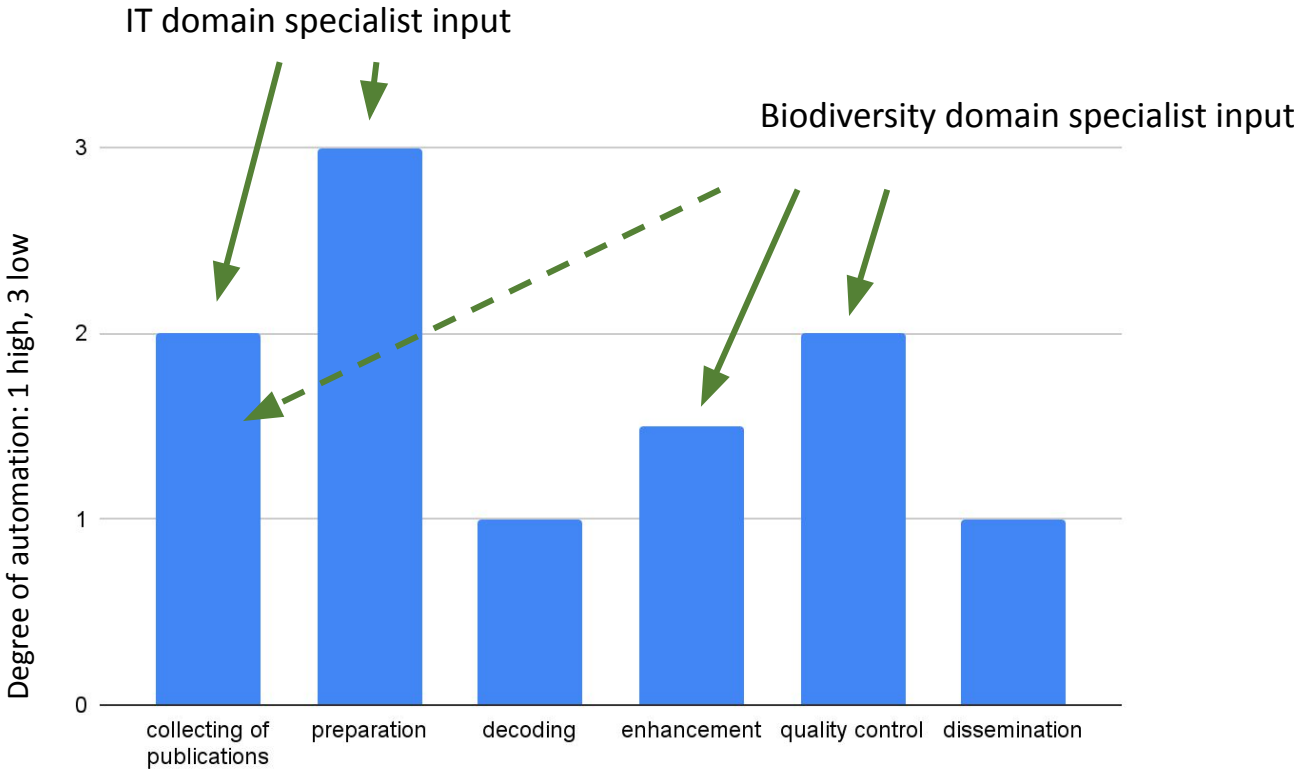
FAIR Digital Objects

Reuse

Research Data Life Cycle



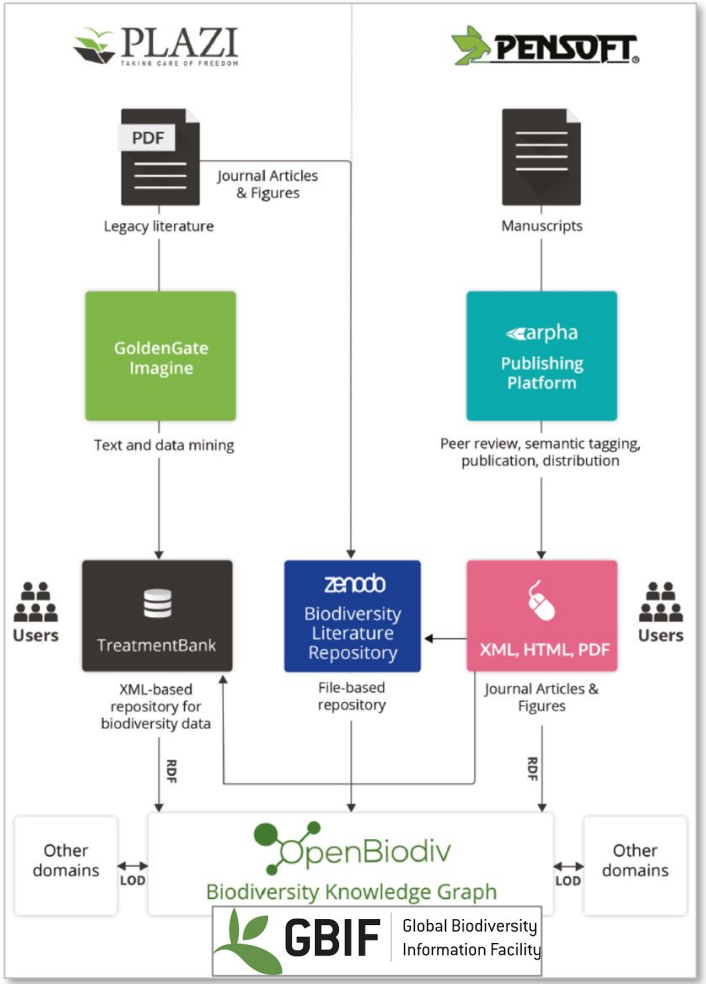
Degree of automation / human input



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.

The future will be open, semantically enhanced publishing

Legacy publications



Prospective publishing

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

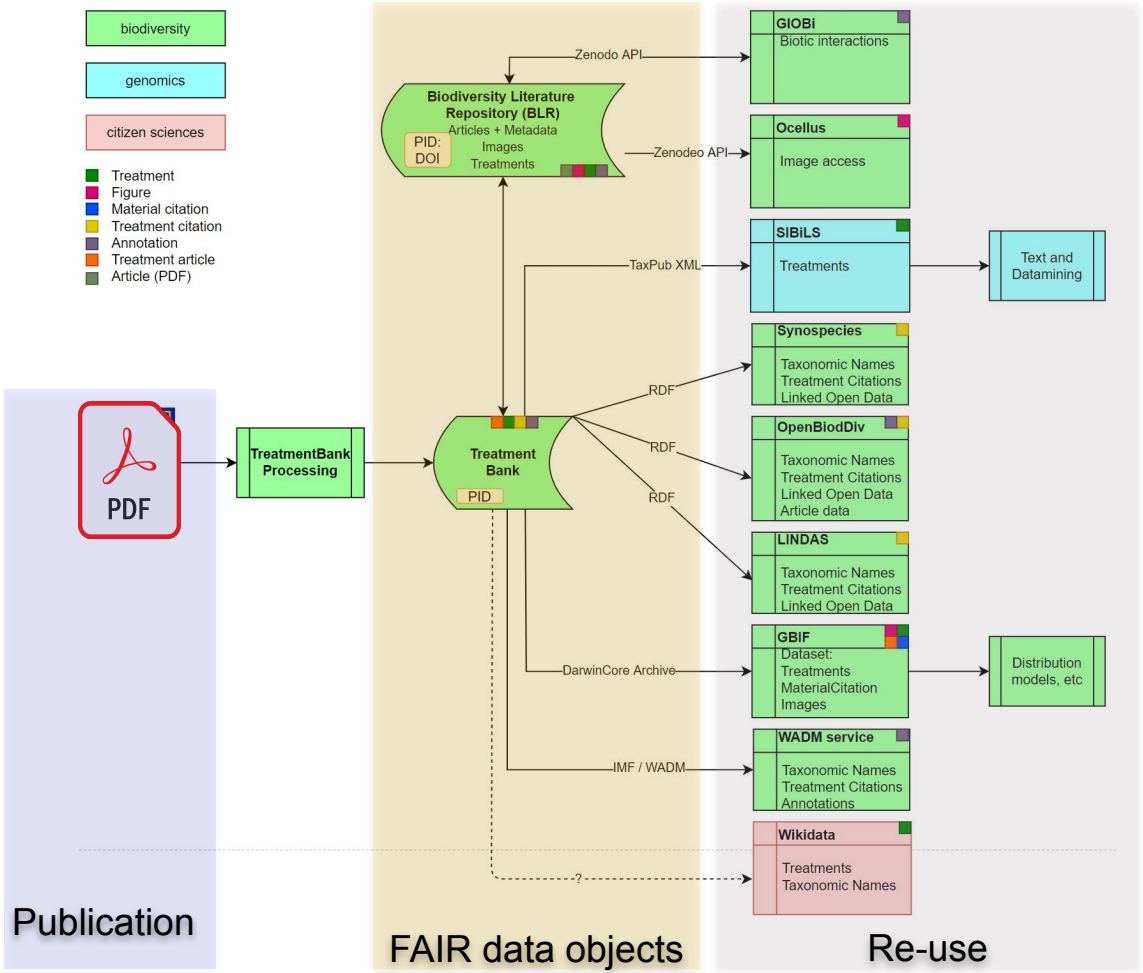
Biodiversity Literature Repository

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

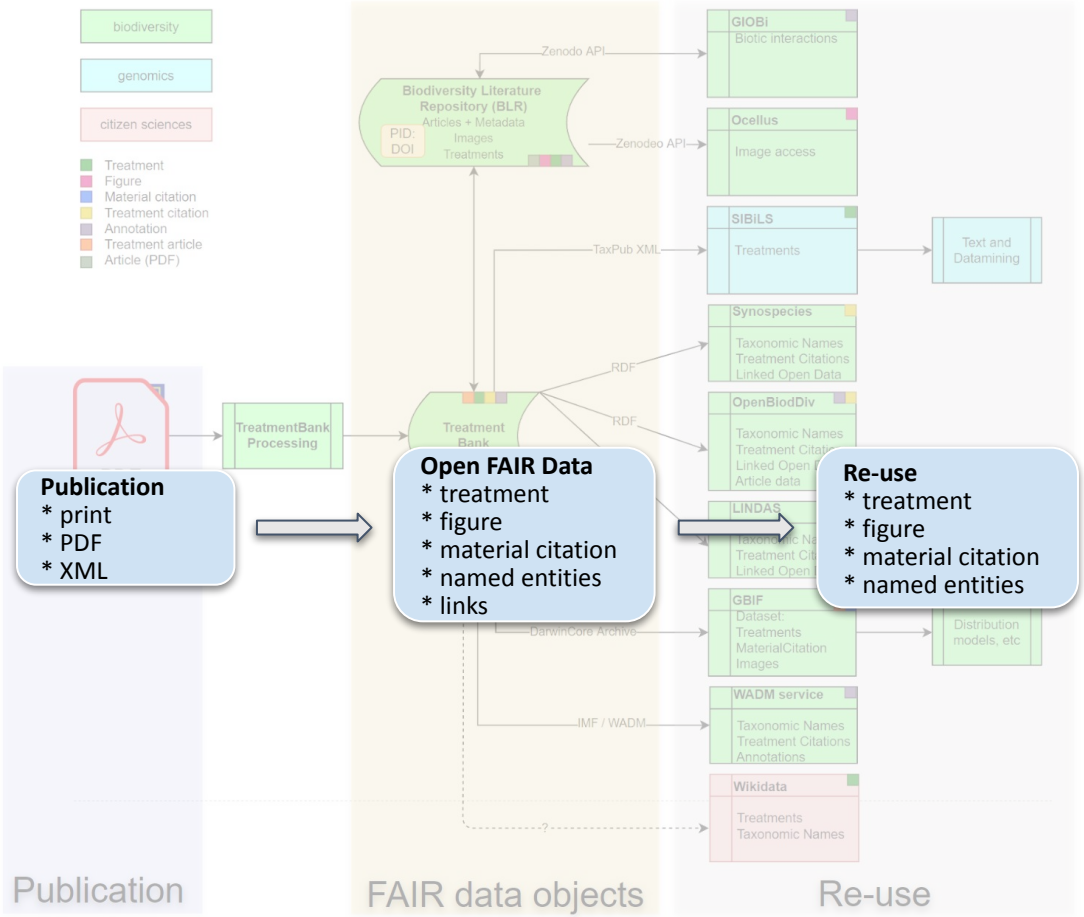
Reuse of treatment articles mediated by Plazi
39,700 treatment article data sets (58% of total data sets in GBIF)
377,000 taxonomic treatments (90,000 unique species)
224,000 figures
591,000 materials citations (occurrences)
The figures are lower in GBIF because of Quality Control measures

Re-use of scholarly publication's FAIR data



FAIR reconsidered

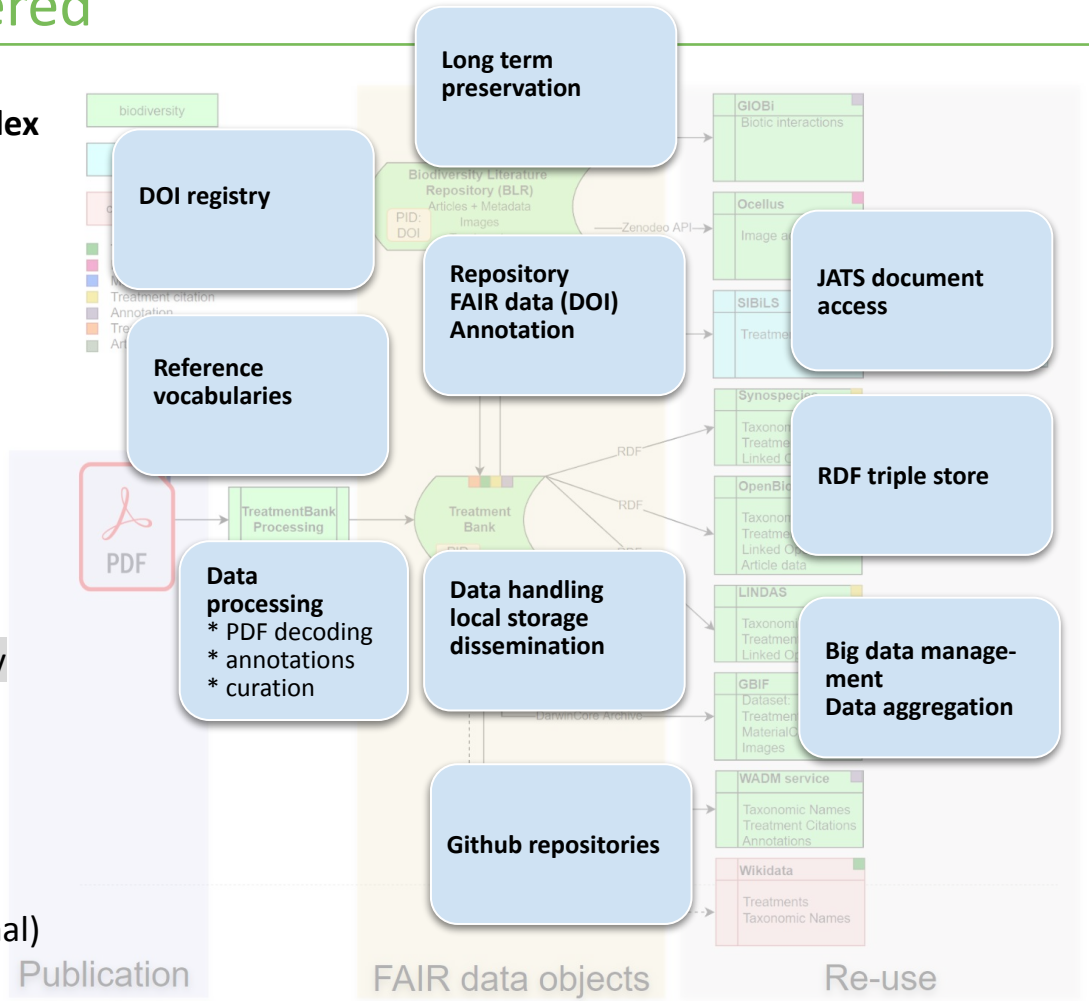
Data flow



FAIR reconsidered

Infrastructures are complex

Each infrastructure needs
* idea
* use case
* initiation
* planning
* management plan
* data policy, licencies
* negotiation
* funding to build
* funding long term
* bilateral interoperability
* reference vocabularies
* user interfaces
* helpdesk
* teaching material
* maintenance
* legal framework (external)



Open FAIR Data

* represent knowledge about biodiversity, i.e.g what we know about a gene, specimen, species as well as assess a gene, specimen in its scientific knowledge, and allows exploring by machine

* is basis for conservation, biodiversity research, society, CBD's post 2020 Global Biodiversity Framework

* depends on complex infrastructures, funding, collaborations producing FAIR data that

* can be used anywhere, at any time by anyone by human or machine.

* increased by accelerating switching to advanced semantically enhanced publishing

* a complex process in need of teaching how to use the results and create innovations

*Access and Benefit Sharing (ABS): **Open FAIR access as part of benefit sharing**

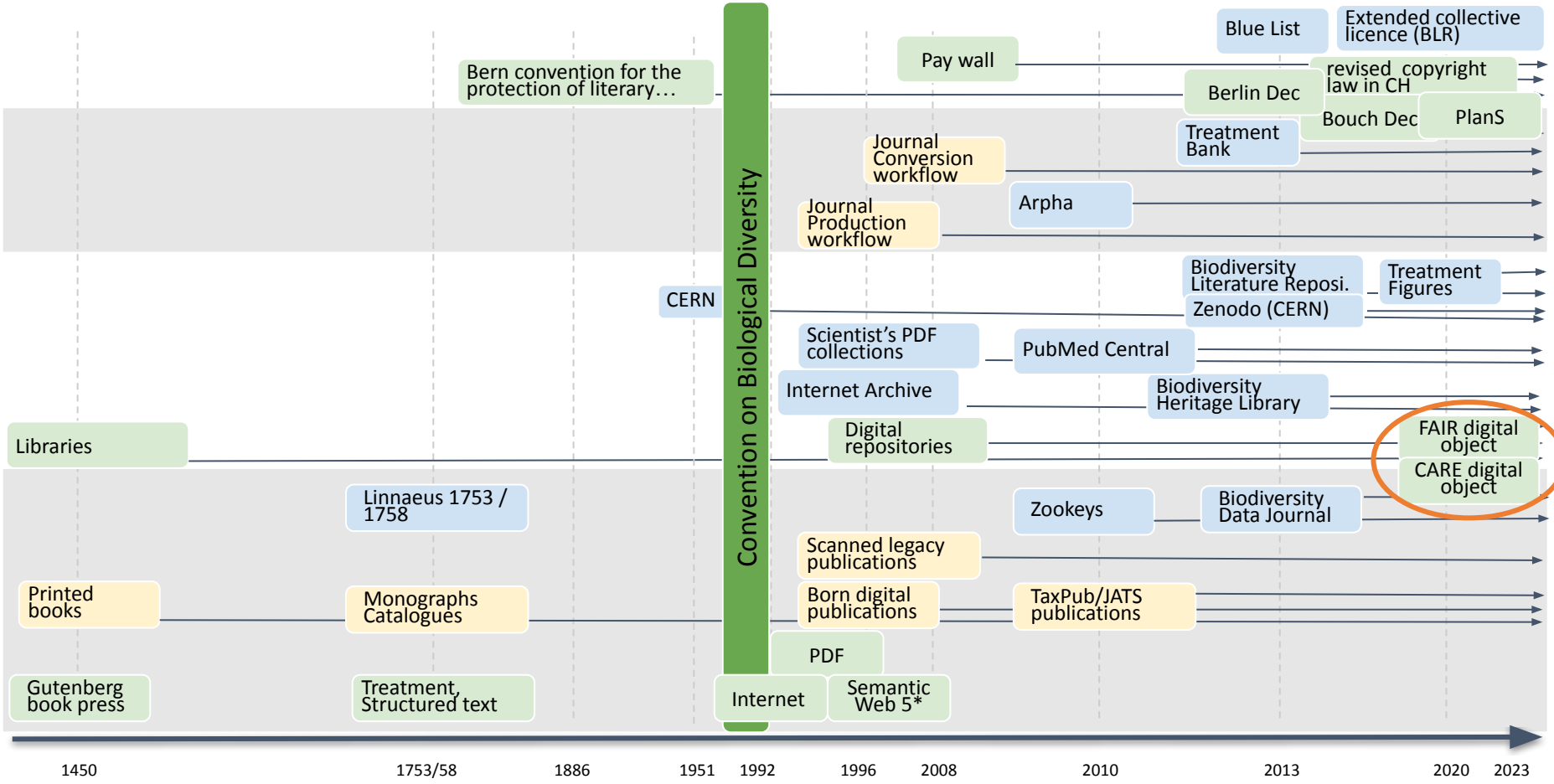
Thank you!

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

Introduction to digitizing taxonomic literature with Plazi DOI

Donat Agosti agosti@plazi.org

Brief history of publishing in biodiversity / taxonomy



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).