A workflow for bidirectional linking of data from literature to external resources

Deliverable D6.3

31 October 2022

Authors

Donat Agosti*, Terry Catapano*, Reto Gmür*, Puneet Kishor*, Conny Naseband*, Guido Sautter*, Alexandre Flament#, Emilie Pasche#, Patrick Ruch#, Alexandros Ioannidis-Pantopikos%, and Jose Benito Gonzalez Lopez %

** Plazi, Bern, Switzerland*
 # Swiss Institute of Bioinformatics / HESGE, Carouge, Switzerland
% CERN, Meyrin, Switzerland

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Start of the project:	May 2021
Duration:	36 months
	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	A workflow for bidirectional linking of data from literature to external resources
Deliverable n°:	D6.3
Nature of the deliverable:	Other
Dissemination level:	Public
WP responsible:	WP6
Lead beneficiary:	Plazi
Citation:	Agosti, D., Catapano, T., Gmür, R., Kishor, P., Naseband, C., Sautter, G., Flament, A., Pasche, E., Ruch, P. , Ioannidis, A., & Lopez, J. (2022). *A workflow for bidirectional linking of data from literature to external resources*. D6.3 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 18
Actual submission date:	31 October 2022

Deliverable status:

Version	**Status**	**Date**	**Author(s)**
1.0	Draft	14 October 2022	Plazi
2.0	Draft	29 October 2022	Plazi
3.0	Draft	29 October 2022	SIB, CERN
3.0	External review	29 October 2022	Tim Robertson, GBIF
4.0	External review	31 October 2022	Lyubomir Penev, Pensoft
5.0	Final	31 October 2022	Plazi, all

Table of contents

Preface

BiCIKL aims at establishing networks of linked data from scientific literature, molecular biology, natural history collections, and taxonomy. From the point of view of literature, this requires unstructured publications to be annotated in a format that represents the target data, for example, accession codes for gene sequences, specimens and taxonomic names, and attributed with the respective persistent identifiers of the target objects.

The requisite annotation is achieved by liberating and FAIRifying the data such as taxonomic treatments, figures, and tables, from publications with the help of the TreatmenBank services in combination with the Biodiversity Literature Repository community on Zenodo (Task 6.2). This allows citation of the liberated data as well as to contribute and complement it as data in infrastructures that are not yet digitally available. While molecular data is born digital and made citable from the beginning, specimen data from natural history collections and taxonomic names have to be digitised retrospectively, after the momentum of their appearance in a collection or in the published literature.. Publishing specimens with a Persistent Digital Identifier (PID) upfront when collected is a very recent practice. Furthermore, digitization of specimens in natural history collections can lag behind material citations in publications, which are the only evidence of a specimen's presence. Publishing this data from natural history collections and scientific literature in the Global Biodiversity Information Facility (GBIF) provides a unique chance to link specimens with their material citations from the literature, and thus, provide direct evidence of the use of a specimen. A material citation is also an entry point to the knowledge graph connecting the specimen from the respective publication to its cited resources therein.

The challenge for TreatmentBank, in the context of BiCIKL, is to enable the Research Infrastructures in the consortium to discover relevant material citations and taxonomic treatments so they can be linked bidirectionally, and to update infrastructures in case those data are not available now but may exist in the literature.

Summary

In this deliverable we report a functional workflow to bidirectionally link cited genomic, specimen and taxonomic data, liberated and FAIR-ized from semantically unstructured publications (T6.2) with the respective research infrastructures that aggregate and host this data. This includes the European Nucleotide Archive (ENA) for genomic data, Catalogue of Life for taxonomic names, and GBIF for specimen data as a proxy for natural history collections. In this process, taxonomic names, treatments and material citations are submitted to Catalogue of Life and GBIF respectively and the identifiers of the records are retrieved.

A workbench is provided to curate links via the GoldenGate Imagine software. A dedicated matching service developed in-house predicts links and also allows users to curate links between material citations in literature and their occurrences in GBIF.

TreatmentBank provides access to all the bi-directional links obtained in this process.

This work is based on the recommendations developed in Task 6.3. to semantically enhance publications with annotations and persistent identifiers to facilitate automated text and data mining workflows (Agosti et al, 2022, in press).

Next steps are discussed such as the reprocessing of the treatments in TreatmentBank to update all the links with the recently established new linking tools with ENA and identifiers for taxonomic names in COL.

List of abbreviations

API	Application programming interface
BLR	Biodiversity Literature Repository
CLB	ChecklistBank
COL+	Catalogue of Life
EU	European Union
ENA	European Nucleotide Archive
FAIR	Findable, Accessible, Interoperable and Reusable
GBIF	Global Biodiversity Information Facility
GUI	Graphical User Interface
JATS	Journal Article Tag Suite
NCBI	National Centre for Biotechnology Information
PDF	Portable Document Format
PID	Persistent identifier
SAH	Source Annotation Helper
SIB	Swiss Institute of Bioinformatics
SIBiLS	Swiss Institute of Bioinformatics Library System
TaxPub	Name of an XML schema used to annotate literature
TB	TreatmentBank
XHTML	eXtensible HyperText Markup Language
XSLT	eXtensible Stylesheet Language Transformations

1. Background and objectives

Taxonomic literature includes an estimated, continually growing corpus of 500 Million published pages that constitute the authoritative knowledge about the world's biodiversity (Kalfatovich, 2010). All known biological species have at least one taxonomic treatment – a well defined section of a scientific publication – where the discovery of a new species is recorded, a new taxonomic name is made available from a nomenclature point of view and a link to a specimen as the holotype is established. Increasingly, publications include material citations of the specimens researched, as well as citations of DNA sequences produced. In subsequent publications, new results are published about known species citing previous treatments, or the synonymy of one taxon with another is established, widening the understanding of biological diversity.

A unique contribution of literature is that it provides the entire history of taxonomic names as a source for the catalogue of life and the links between treatments and specimens, and more recently, also DNA sequences. In reverse, this offers to explore what is known about a specimen or a gene sequence.

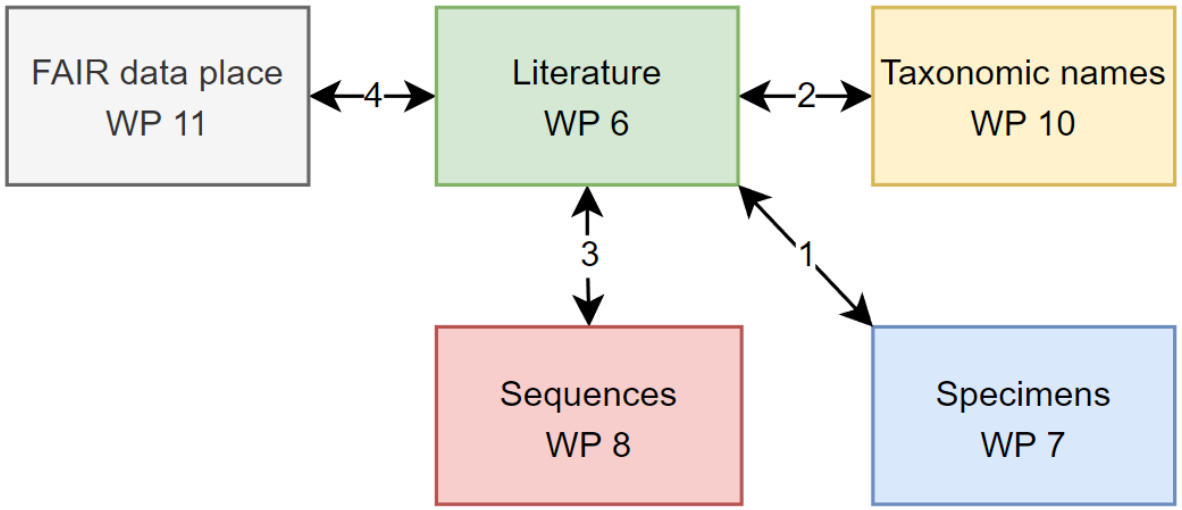

Figure 1. *Overview of the data domains represented in BiCIKL and the bidirectional links (1-3) produced from literature, and access to the links from the FAIR data place (4). In BiCIKL, each domain is linked to each other as well as directly to the FAIR data place.*

The challenge is to make the pertinent data findable, accessible, interoperable and reusable (FAIR) and link it with the cited resources. While task 6.2 is providing the data conversion workflow (Agosti & Egloff, 2008), the objective of this task is to bidirectionally link data from literature to external resources. Specifically, this involves defining a workflow that provides a formal description, transfer protocols and exchange mechanisms to automate the linking, as well as manually curating the individual links.

2. Workflow

2.1. Data

The data listed in Appendix 1 are annotations made in the data conversion process (T6.2) in the unstructured publications. The annotated document is available if the publication is open access, otherwise the respective treatments only are made available. The latter, the focus in BiCIKL, are complemented with figures, tables, and material citations. Treatments and figures are converted into open FAIR data in BLR (Agosti & Ioannidis-Pantopikos, 2022). Tables are extracted and made accessible through TB[1], and citable with a persistent identifier[2]. Because of open questions on the quality control, they are not yet deposited in BLR. Material citations are accessible as citable data in GBIF[3]. The main annotated entities include specimen, accession and collection codes, as well as geographic coordinates, dates, and person names.

2.2. Transfer formats

Different data transfer formats exist because a generic format is required to represent all the different aspects (GG XML), and the recipients are using more specific formats, often designed to represent only some aspects of the data: Darwin Core Archive (DwC-A) for the import for GBIF, TaxPub/JATS XML for SIBiLS, or JSON/XHTML to export data to BLR/Zenodo.

The respective files are produced in real time and readily available, or a recipient system such as GBIF is notified of updates and it harvests the data later.

[1] https://tb.plazi.org/GgServer/tpsStats/
[2] http://table.plazi.org/id/UUID see e.g.
http://table.plazi.org/id/9B4F6833A2C2B7B7ACBB815A130D9B3A or in JSON
http://table.plazi.org/id/9B4F6833A2C2B7B7ACBB815A130D9B3A?format=JSON
[3] https://www.gbif.org/occurrence/3764007302

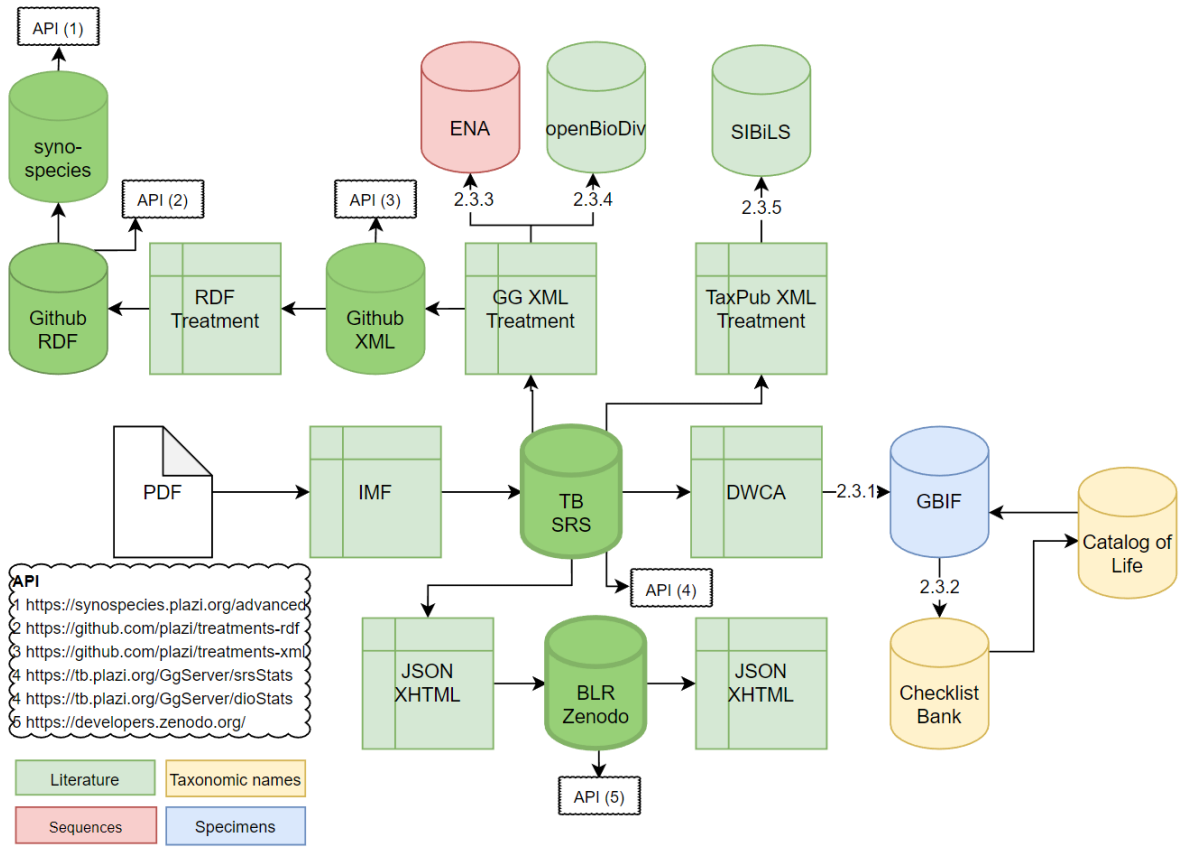

Figure 2. *Overview of the data flow between literature based data and the data domains in BiCIKL (specimens, taxonomic names, and sequences).*

2.2.1. Darwin Core Archive (DwC-A)

The Darwin Core Archive format is a simple and extensible arrangement of tabular data in a star schema for sharing biodiversity data, especially catalogue data based on the ratified Darwin Core terms and the Darwin Core text guidelines (GBIF, 2021; Smirnova et al. 2016). An outline is available at TreatmentBank[4].

2.2.2. TaxPub XML (Treatment)

JATS TaxPub is used to import treatments into SIBiLS. JATS is actually maintained as standard by a NISO committee[5] also the data format used by PMC, the full text archive of the world's biomedical literature, and EuropePMC, the European mirror. Enabled by the JATS, SIBiLS annotations are copied into EuropePMC via the SciLite gateway (Venkatesan et al. 2017). Similarly, we plan to also copy biodiversity specific annotations (e.g. biotic interactions) of SIBiLS within EuropePMC - see D11.3. The upload of taxonomic treatments into other repositories such as EuropePMC is also a work in progress. Furthermore, the definition of the TaxPub data levels are provided here[6].

[4] http://plazi.org/treatmentbank/treatment-data-access/#appendix-darwin-core-archive-content
[5] https://www.niso.org/standards-committees/jats
[6] https://github.com/plazi/qqxml2taxpub-treatments/blob/main/taxpub%20levels.md

2.2.3. JSON / XHTML (Treatment, Figures)

JSON format is used to interact with Zenodo. It follows the terminology provided by the API description[7]. XHTML is used as the format for taxonomic treatment as the digital object required to create a Zenodo deposit.

2.2.4. GG XML (Treatment)

Plazi's internal XML format used for storing treatments in TreatmentBank, and as the basis for transformation in all the other formats listed in this section, represents all the significant features and structures of treatments such as nomenclature acts, treatment citations, materials citations, citations of figures, tables, and other publications, accession and barcode numbers, etc. Representation of basic treatment structures is enforced by rules akin to the XPATH based Schematron validation, as are the markup patterns representing features such as treatment citations and materials citations; the intertwining of other textual components and patterns, on the other hand, is more flexible to allow reflecting the semantics of any given treatment with respective markup (see Appendix 1 for a complete list of annotation types / element names used). This open-world approach also facilitates integration of possible markup of other aspects without dependency on the evolution of an XML schema.

2.3. Exchange mechanisms

2.3.1. Specimens (GBIF)

Data transfer to GBIF uses DwC-A (see 2.2.1), with each archive bundling the data of all treatments from an individual source publication. The archives are (re-)generated as new treatments are added/marked in a publication, or existing ones are modified. After an archive is packed, it is either registered to the GBIF API (on first export), or the GBIF API gets notified of the update (on subsequent exports). GBIF then downloads and ingests the DwCA from TreatmentBank.

2.3.2. Taxonomic Names (Catalogue of Life)

Newly created taxon names, as well as other nomenclature acts, are included in the DwCAs exported from TreatmentBank (see 2.2.1 and 2.3.1), and thus ingested by GBIF. From there, the newly coined taxon names get forwarded to Catalogue of Life.

2.3.3. DNA sequence data (ENA)

Accession numbers cited in treatments are marked as such and linked to their respective sequence pages in ENA. In the other direction, ENA follows an update feed that informs it about new treatments as well as modifications to existing ones. Based upon this feed, ENA

[7] https://developers.zenodo.org/

discovers treatments and material citations associated with the names of the taxa from whose specimens gene sequences are derived.

2.3.4. Linked Open Data (OpenBioDiv)

As new treatments are added to TreatmentBank and existing ones are modified, each one is registered in OpenBioDiv and enqueued for processing. OpenBioDiv then fetches the generic GG XML (see 2.2.2) via the registered HTTP URI and ingests it into its knowledge base.

2.3.5. Treatment TaxPub (SIBiLS)

As new treatments are added to TreatmentBank and existing ones are modified, each one is transformed into TaxPub via XSLT, validated, and pushed to SIBiLS via SFTP.

2.3.6. XHTML (BLR/Zenodo)

As new treatments are added to TreatmentBank and existing ones are modified, each one is transformed into XHTML (see 2.2.3) via XSLT, validated, and pushed to Zenodo via their API. The associated metadata is sent along as JSON, generated from the metadata of the source publication as well as details extracted from the treatment proper. The returned deposition number and the derived DOI are stored back into the treatments.

For newly added publications, a similar mechanism exports the underlying source PDF to Zenodo and stores the returned deposition number in the converted publication; the derived DOI is stored only if the source publication does not come with a DOI that was minted and assigned by the publisher.

Furthermore, individual figures and graphics are exported to their own individual Zenodo depositions as well (as PNGs), and the returned deposition numbers and derived DOIs are stored in their associated captions in the converted publications. The DOIs are also added to in-text citations of the figures, thereby establishing the link between treatments and the figures they cite.

2.4. Workflow

2.4.1. Specimen (GBIF)

Upload
GBIF/ChecklistBank: a DwC-A is packed and the GBIF API is notified about the update, GBIF then downloads and ingests the DwC-A into both the specimen handling systems (GBIF.org) and the taxonomic databases (Checklistbank.org).

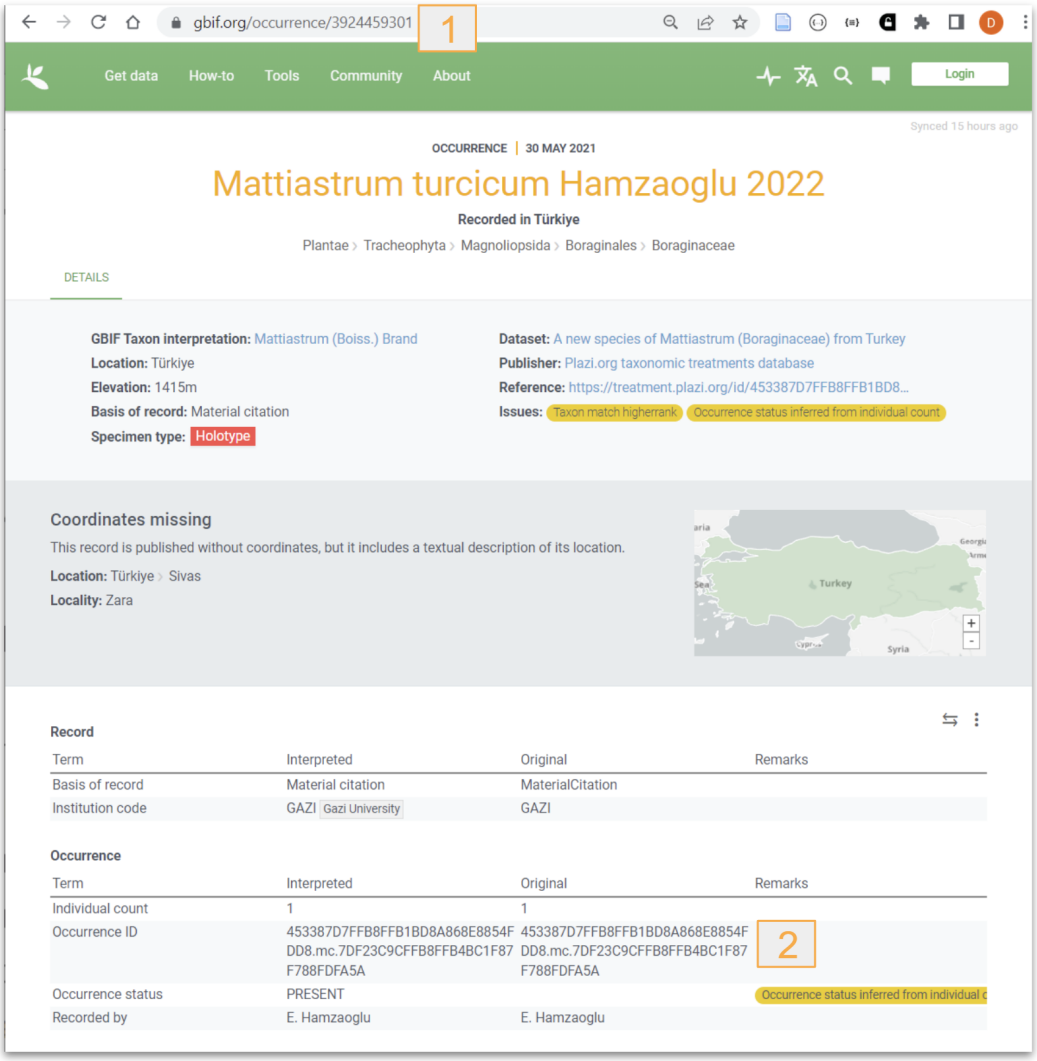

Figure 3: *Material citations re-used by GBIF. 1: GBIF occurrence key; 2: Imported TB occurrence ID.*

Import of identifiers

The GBIF API returns a dataset key when a new dataset is registered, and TreatmentBank stores that dataset key in the source publication.

After (re-)exporting a DwC-A and notifying GBIF, a 15 minute timer starts, at the end of whichTreatmentBank fetches the records via the dataset key, extracts the keys of the individual taxon and occurrence records, and adds them to the respective treatment taxa and material citations; if nothing is found, the timer resets to wait for another 15 minutes and tries again.

Treatment UUID	Verbatim Taxon Name	Materials Citation UUID	GBIF Occurrence ID
453387D7FFB8FFB1BD8A868E8854FDD8	Mattiastrum turcicum Hamzaoğlu	7DF23C9CFFB8FFB4BC1F87F788FDFA5A	3924459301

Figure 4: *Bidirectional links between TB (Materials Citation UUID) and GBIF material citation (GBIF Occurrence ID)*[8].

8

https://tb.plazi.org/GgServer/srsStats/stats?outputFields=doc.uuid+tax.name+matCit.id+matCit.gbifOccur

2.4.2. Taxonomic Names (Catalogue of Life)

Upload

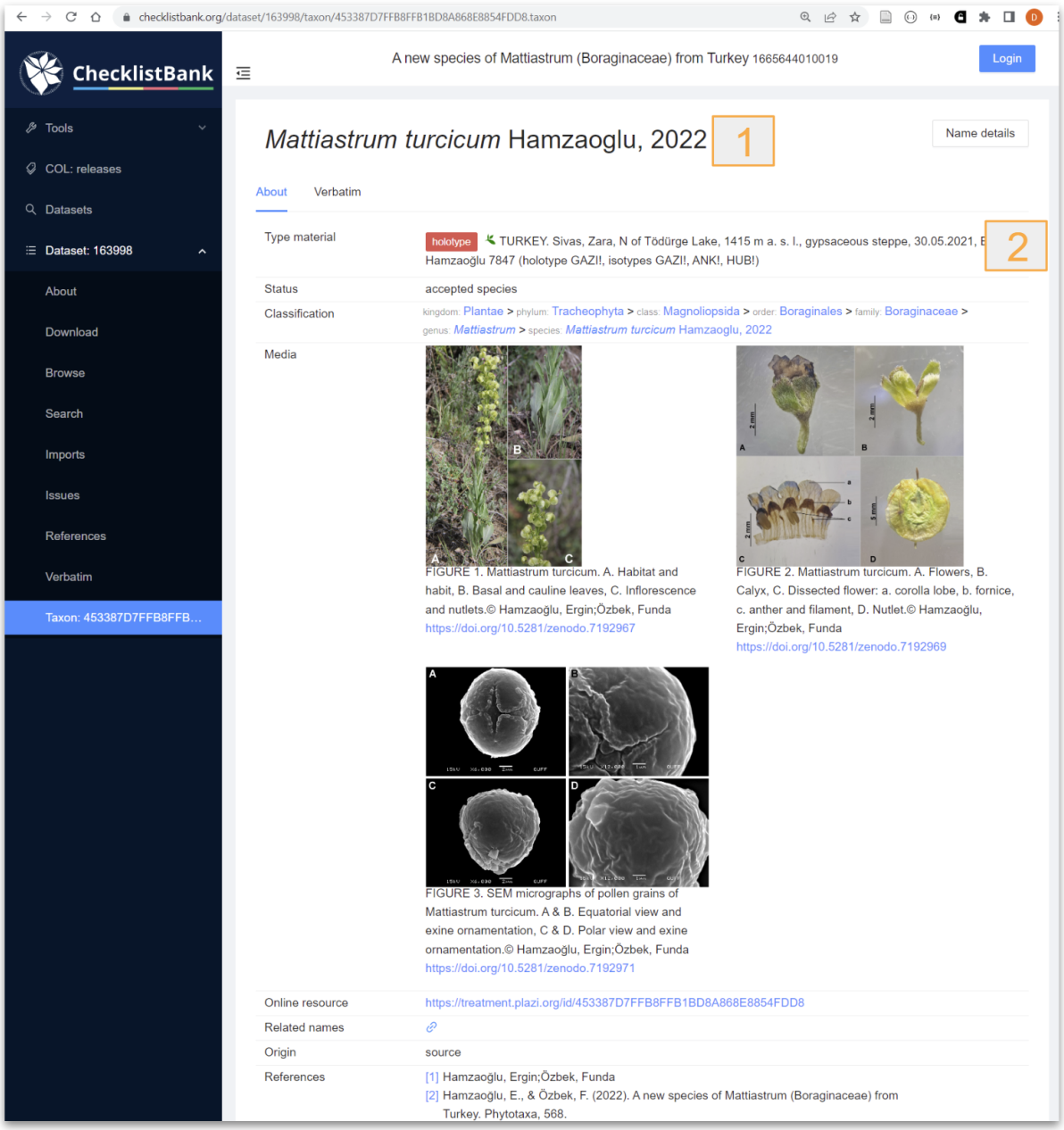

Figure 5. *Data submitted and reused from ChecklistBank[9]. 1- taxonomic name, 2 Material citation. The taxonID in Checklistbank for the taxonomic name in the treatment is the same as the one minted by TB.*

Bidirectional Linking

For new taxon names, this will be a timer-based approach, checking for the newly minted taxon name identifiers 6-12 hours after a recent original description has been exported to GBIF via a DwC-A.

---

renceId&groupingFields=doc.uuid+tax.name+matCit.id+matCit.gbifOccurrenceId&FP-matCit.gbifOccurrenceId=3924459301&format=JSON

[9] source: https://www.checklistbank.org/dataset/163998/taxon/453387D7FFB8FFB1BD8A868E8854FDD8.taxon

For existing taxon names, the respective identifiers from Catalogue of Life will be added as part of the original treatment extraction process.

2.4.3. DNA sequence data (ENA)

As part of the markup process, TreatmentBank will do a lookup to ChecklistBank to get taxon name identifiers from Catalogue of Life and their associated identifiers from the NCBI/ENA taxonomic backbone, and then add these identifiers to the treatment taxa.

Annotated accession numbers will be attributed with the respective HTTP URI, pointing to the ENA information page for the cited gene sequence. In this case, no lookups are required, as the accession numbers proper are the sequence identifiers to use in the HTTP URIs.

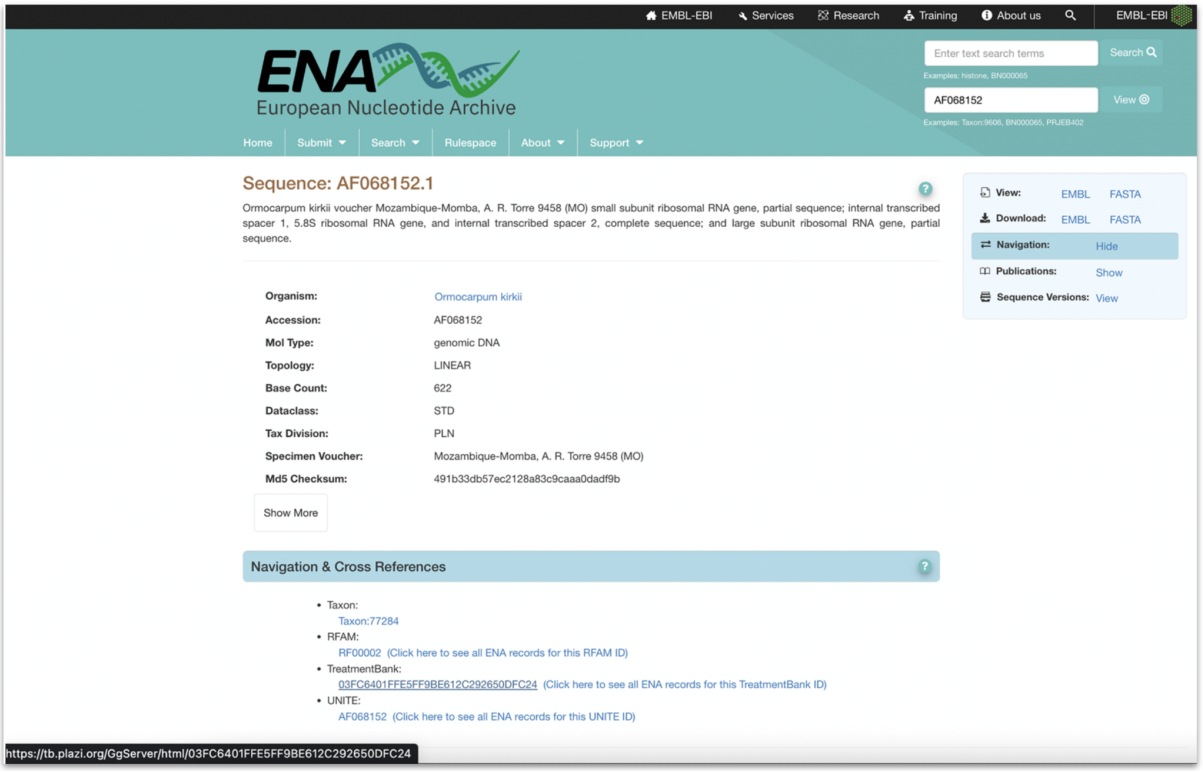

Figure 6: *ENA Mockup display of cross reference to TreatmentBank.*

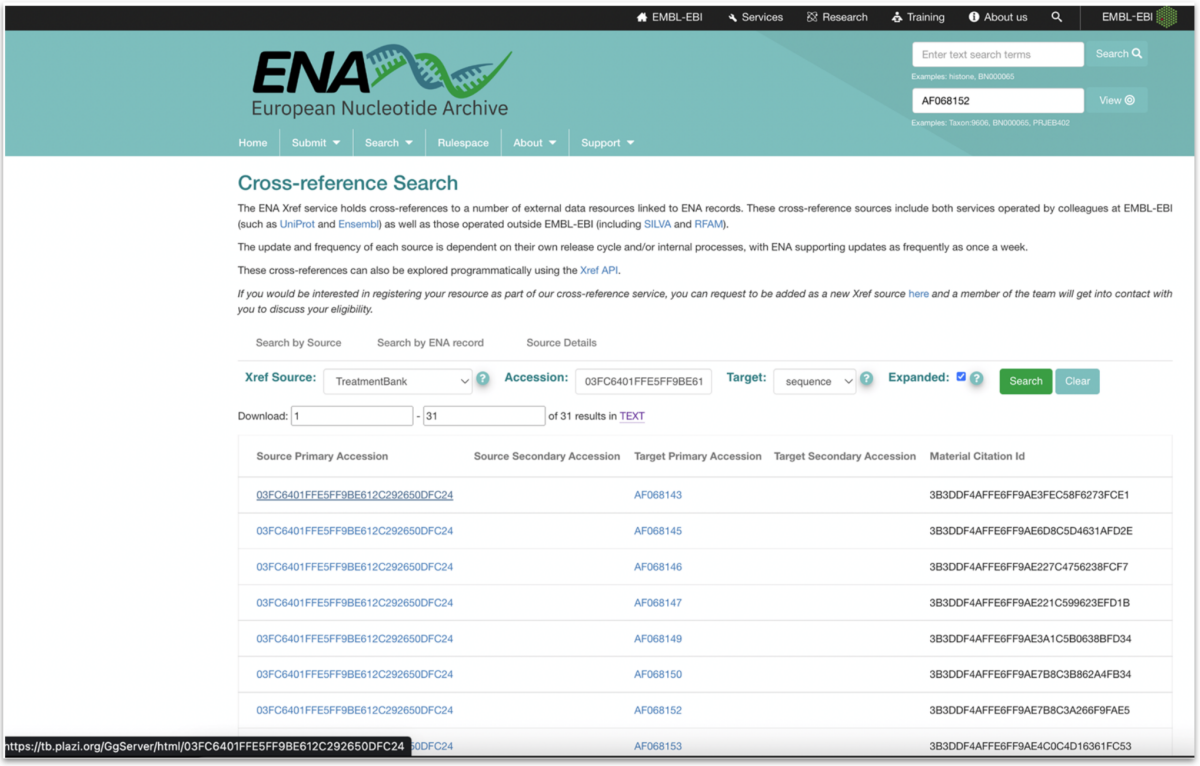

Figure 7: *Mockup of Cross-reference Search for accession numbers in publications via TB. Links to treatments (Source Primary Accession) and material citation (Material Citation ID).*

2.4.4. Linked Open Data (OpenBioDiv)

As new treatments are added to TreatmentBank and existing ones are modified, each one is registered to OpenBioDiv and enqueued for processing there. OpenBioDiv then fetches the generic GG XML (see 2.2.2) via the registered HTTP URI and ingests it into its knowledge base.

2.4.5. Treatment TaxPub (SIBiLS)

As new treatments are added to TreatmentBank and existing ones are modified, each one is transformed into TaxPub via XSLT, validated, and pushed to SIBiLS via SFTP.

Bidirectional links to SIBiLS will be generated once the SIBiLS Graphic User Interface is made public.

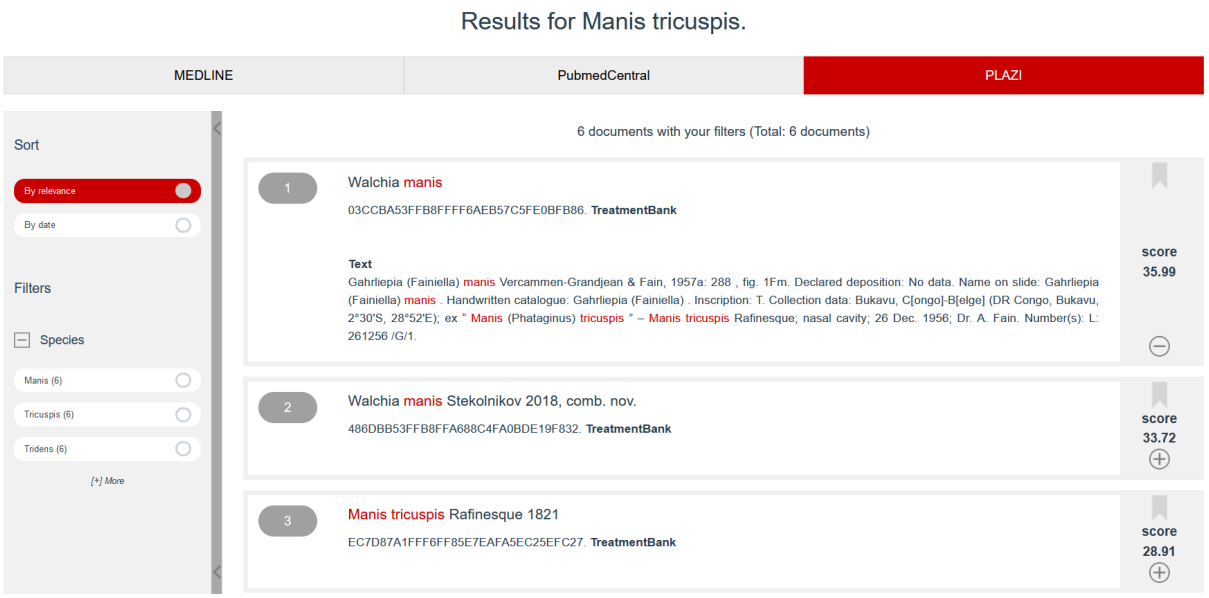

Figure 8: *Display of Plazi contents in SIBilS In this example, the Plazi unique TreatmentBank identifier is a hyperlink, which allows the reader to directly navigate back to Plazi.*

2.4.6. Linked Open Data (Synospecies)

As new treatments are added to TreatmentBank and existing ones are modified, updates are bundled by both time and sheer number, and the generic GG XML is pushed to a dedicated GitHub repository. A GitHub workflow, triggered by the push, picks them up and transforms the generic GG XML[10] into RDF XML[11] and Turtle, which is subsequently ingested into Synospecies.

2.4.7. XHTML (BLR/Zenodo)

As new treatments are added to TreatmentBank and existing ones are modified, each one is transformed into XHTML (see 2.2.3) via XSLT, validated, and pushed to Zenodo via their API. The associated metadata is sent along in JSON, generated from the metadata of the source publication as well as details extracted from the treatment proper.

The returned deposition number and the derived DOI are stored back into the treatments.

Alongside the treatments, TreatmentBank also exports the source PDFs of newly added publications to Zenodo via a similar mechanism and stores the returned deposition number in the converted publication; the derived DOI is stored only if the source publication does not come with a DOI that was minted and assigned by the publisher.

Furthermore, individual figures and graphics are exported to their own individual Zenodo depositions as well (as PNGs), and the returned deposition numbers and derived DOIs are stored in their associated captions of the converted publications. The DOIs are also added to

[10] https://github.com/plazi/treatments-xml
[11] https://github.com/plazi/treatments-rdf

in-text citations of the figures, thereby establishing the link between treatments and the figures they cite.

3. Workbench

Data automatically liberated from literature (T6.2) can be annotated using GoldenGATE Imagine. This also allows to curate and attribute annotations like “accession code”, “collection code”, “specimen code” or “taxonomic name” with the respective identifiers.

The Matching service is a dedicated tool to link material citations to preserved specimens in GBIF. For this purpose, GBIF is providing all the clusters of record relations it detects, including material citations, to TreatmentBank where, for each material citation, one or more matches are provided.

3.1. GoldenGATE Imagine interface

The GoldenGATE Document Editor is a visual editor for marking up documents. The main goal of the Imagine version is to provide a markup tool for digital born PDFs. It is designed to do most of the markup automatically; manual work is reduced to correcting the output of automated components[12]. Accession numbers, specimen codes or collection codes can automatically be detected and attributed with the respective PIDs. However, there is no standard way to publish and compose these codes or numbers yet, and thus automated annotation needs human curation, which is enabled using the GGI UI (Fig. 9).

Access to new documents or editing existing annotations requires authentication depending upon training[13], which will be provided in collaboration with WP3. A web based version of the material citation UI will be provided in the second part of BiCIKL. Enhanced publications are versioned and can be rolled back to previous versions and assigned to a specific user (Fig 10).

[12] http://plazi.org/treatmentbank/desktop-data-mining/
[13] https://github.com/plazi/community

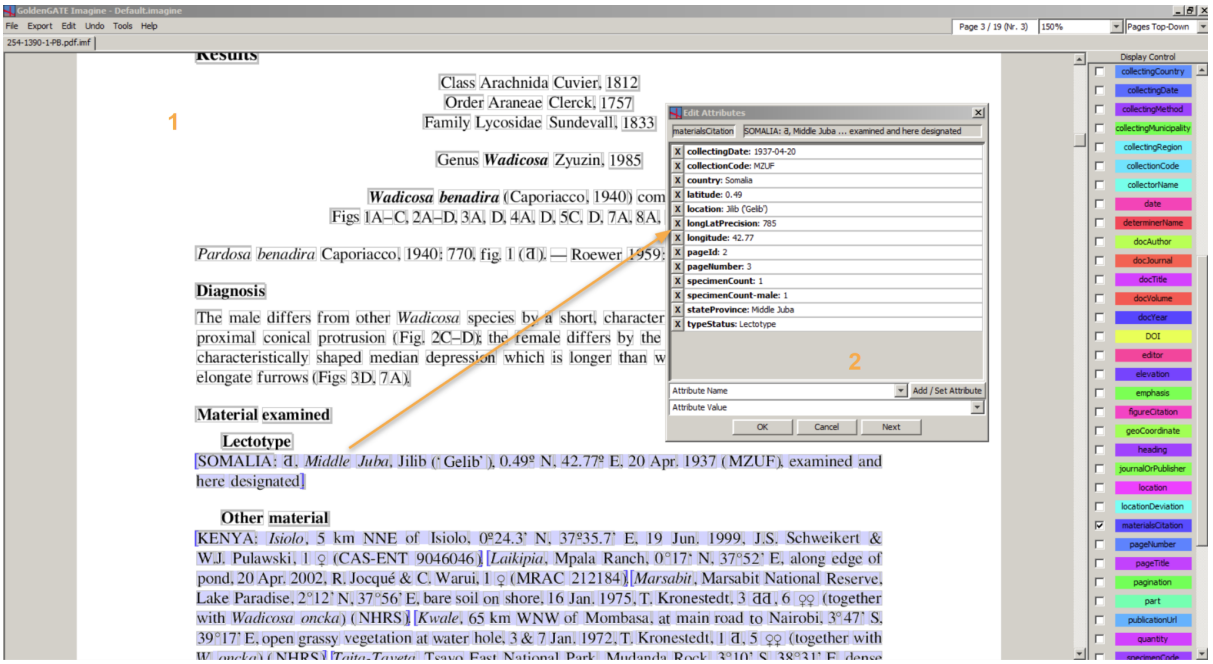

Figure 9: *GoldenGate Imagine editor User Interface. 1 Main panel; 2 annotation editor.*

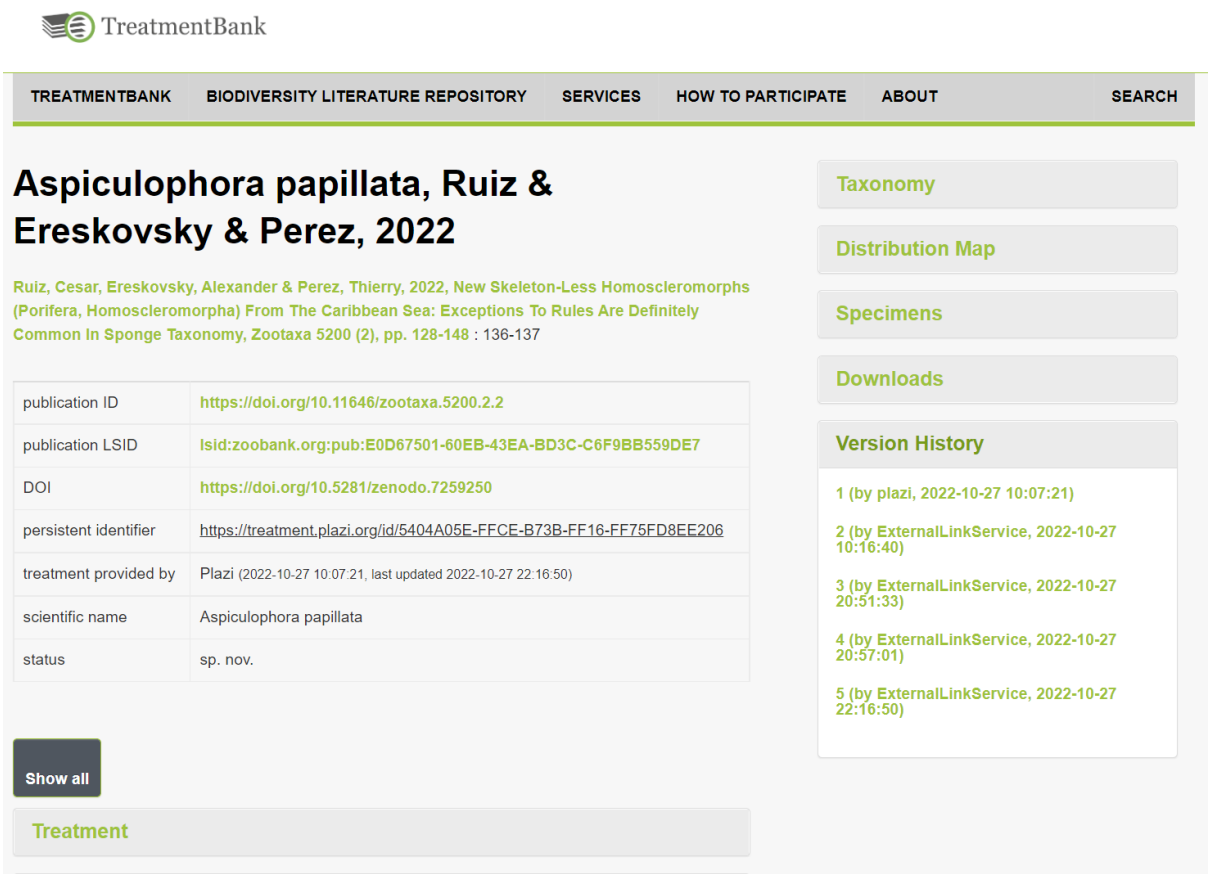

Figure 10: *Treatment view in TreatmentBank showing the version history*[14].

[14] https://tb.plazi.org/GgServer/html//5404A05EFFCEB73BFF16FF75FD8EE206

3.2. Matching service

Material citations in publications traditionally cite a single preserved specimen, such as a holotype or groups of specimens. Material citations as part of a treatment provide scientific results based on the respective specimen, and in other words, add the specimen to the biodiversity knowledge graph.

Today, most of the digital specimens are not accessible via APIs at their natural history museums, but the datasets are often published in GBIF. This provides an opportunity to use a single consistent matching service instead of developing individual solutions.

The Matching service (Fig. 11) is based on the clustered occurrence data in GBIF, namely clusters including material citations from Plazi as well as occurrence records that represent museum specimens (Meeus et al., 2022). In reality, the data have different codes, abbreviations, and omissions, so detailed match-up is based upon a one-by-one comparison of individual attributes of specimens that are cited in material citations, yielding per-attribute and overall matching scores. Once a curator or taxonomer has confirmed a match between a material citation and a museum specimen by inspecting the field comparison in the respective GUI (Fig. 11), that relationship is stored in the matching service internally, which then forwards the GBIF occurrence key of the specimen to TreatmentBank (via a dedicated REST endpoint / microservice), where it is written into the treatments, specifically into the material citation, and forwarded to GBIF when the DwCA is next published. Currently, the link to the GBIF occurrence is recorded in the DwC field “reference” in the occurrence.txt file. As an alternative to the GBIF occurrence key, any PID of a matching specimen can be added.

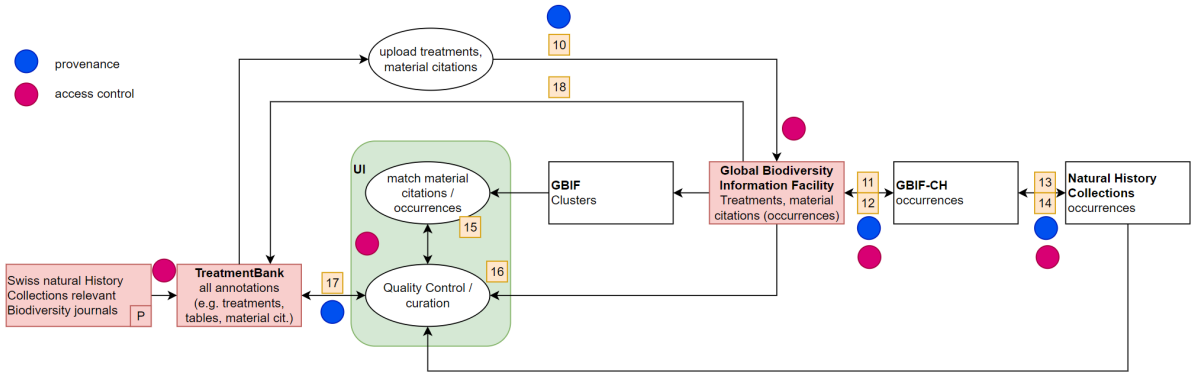

Figure 11: *Schematic Matching service workflow.*

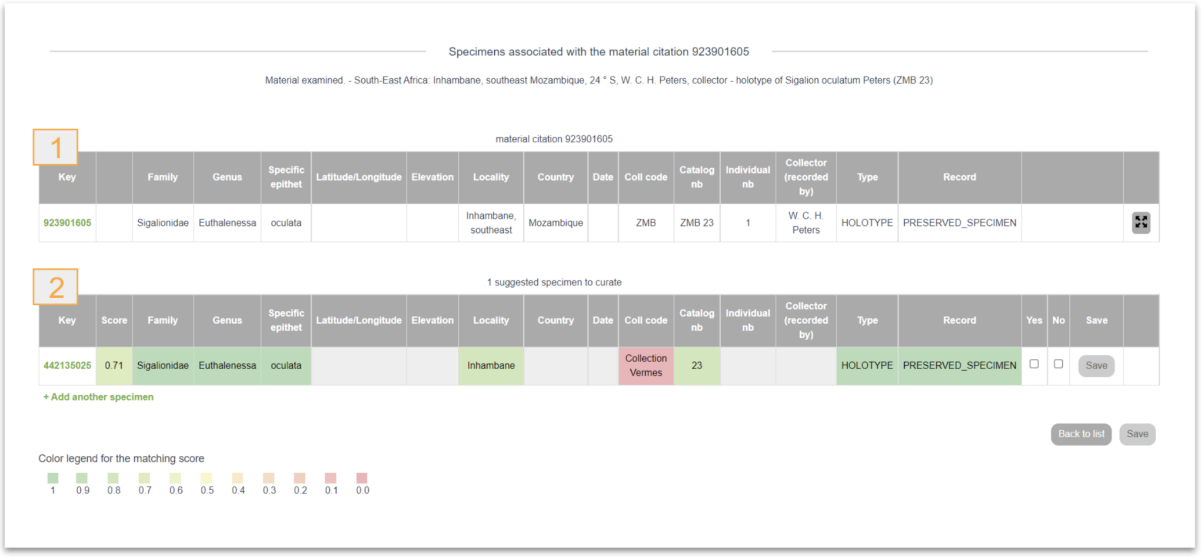

Figure 12: *GUI displaying the match of a material citation reused by GBIF with a preserved specimen. Note the variety in the data fields. 1: material citation GBIF key; 2: preserved specimen GBIF key* [15].

4. Data and source code access

4.1. Data

4.1.1. Data access

The data liberated from the publications is available in open FAIR format from the Biodiversity Literature Repository, and from TreatmentBank directly (SRS stats[16]; introduction[17]).

With GBIF, 630,000 material citations and 137,000 treatments are bidirectionally linked. With ChecklistBank 41,000 checklists are linked, and 690,000 taxonomic names are bidirectionally linked by sharing the same identifier (dwc:ID). The exact number of bidirectional links between ENA and TreatmentBank based on the recently established transfer protocol will be available once the treatments in TB are reprocessed with the new accession code tagger based on the collaboration with ENA, further augmented with the use of data within tables.

15 https://prod.ebiodiv.org/?institutionKey=d742cdc8-729f-4a22-8f2c-84c4bd9dc11b&datasetKeys=7b8964ea-f762-11e1-a439-00145eb45e9a&format=matcit_specimen#occurrences

16 https://tb.plazi.org/GgServer/srsStats

17 https://doi.org/10.5281/zenodo.7262609

4.1.2. Known data issues

Scientific publications are traditionally made for human consumption, and have limited common structure. This makes finding detailed content and expressing its semantics in respective annotations a highly complex task, that varies across journals. Consequently, a human is needed to curate and apply quality control to the data and links (see workbench description above). However, the growing amount of available data allows us to better understand the structure of accession numbers and specimen codes in use, and thus we can adapt and improve the automated extraction algorithms to increase their recall precision. The reuse of correctly annotated and attributed specimen codes and accession numbers in OpenBioDiv and other knowledge management tools will raise the awareness of how data is published in the future, resulting in guidelines for authors and data publishers.

4.1.3. Bidirectional links

At the moment, all the bilateral links for material citations and treatments used by BiCIKL RI can be obtained from the TreatmentBank statistics. This includes treatment taxa, bibliographic metadata of the source publications, treatment citations, materials citations, accession numbers, basic treatment provenance, and several other aspects.

Treatment UUID	Treatment DOI	GBIF Treatment Taxon ID	Article GBIF Dataset ID	Verbatim Taxon Name	Materials Citation UUID	GBIF Occurrence ID	Specimen Code	Accession Number	Accession HTTP URI
03FF87C7FFD5FFF0FF51FF1EFE711E62	http://doi.org/10.5281/zenodo.7157842	202158276	06f3b611-226b-426b-8526-5337e21f5e1a	Planiliza klunzingeri (Day, 1888)	3B3E3C8CFFD5FFF2FE2BF978FDBF1B2D	3921380327	MNHN 2019-0086	MT999034	http://www.ncbi.nlm.nih.gov/nucleotide/MT999034

Figure 13: *Query response to display the known identifiers for the material citation of Planiliza klunzingeri (Day, 1888) in treatment UUID=03FF87C7FFD5FFF0FF51FF1EFE711E62*[18].

4.2. Source Codes

The source code of the programs are available on github.

5. Future steps

The current development that led to an operational workflow, and the new interdependencies are providing novel ways to assess and improve data quality. These insights will be used to improve the quality of data along the production workflow from promoting the

18 https://tb.plazi.org/GgServer/srsStats/stats?outputFields=doc.uuid+doc.doi+doc.gbifTaxonId+doc.articleGbifId+tax.name+matCit.id+matCit.gbifOccurrenceId+matCit.specimenCode+matCit.accessionNumber+matCit.accessionHttpUri&groupingFields=doc.uuid+doc.doi+doc.gbifTaxonId+doc.articleGbifId+tax.name+matCit.id+matCit.gbifOccurrenceId+matCit.specimenCode+matCit.accessionNumber+matCit.accessionHttpUri&limit=100&FP-tax.name=%22Planiliza%20klunzingeri%20(Day%2C%201888)%22&FP-matCit.gbifOccurrenceId=1-&FP-matCit.accessionNumber=MT999034&FP-matCit.accessionHttpUri=1-&format=JSON

recommendations on annotations in publishing of scientific publications as outcome of T6.3[19] via a roundtable with publishers and editors in December 2022 and followup workshop in May 2024 (WP3), to raise the precision of the annotations for accession and collection codes.

The implementation of the linking mechanisms led to an improvement of the annotation tools. In the followup the entire documents will be reprocessed to find named entities such as accession numbers with higher precision and attribute them with their respective PIDs.

A focus will be the extraction of tables, and finding ways to annotate and attribute these with identifiers and specimen codes, institution codes, and accession numbers within the tables. Tables are playing an increasingly important role as a vehicle to publish semi-structured data and are thus a very rich source for bidirectional linking.

Taxonomic names in treatments will be attributed with Catalogue of Life's persistent COL identifier once the treatments are integrated from ChecklistBank to COL (D10.3).

It should be noted that the work presented here pilots the first curation and quality control workflow that builds on the automated GBIF data clustering output. This provides a solid foundation to guide research communities that have similar challenges in using GBIF clusters.

6. Acknowledgements

Plazi and SIB gratefully acknowledge the colleagues in all the RI in BiCIKL with whom we had the chance to collaborate, to ponder ideas, learn a lot and not least to implement all the bidirectional linking mechanisms. A special thanks to Tim Robertson in helping to get the cluster-data from GBIF and implement material and treatment citations in GBIF, and Markus Döring for the strong collaboration with WP10 and continued generous support in implementing the import of taxonomic names to ChecklistBank, and Joana Pauperio and Vikas Gupta for the discussions regarding sequence data..

7. References

Agosti D, Egloff W 2009. Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2009, [2:53]. doi: 10.1186/1756-0500-2-53

Agosti D, Ioannidis-Pantopikos A 2022. Taxonomic Treatments as Open FAIR Digital Objects. Research Ideas and Outcomes 8: e93709. doi: 10.3897/rio.8.e93709

Agosti D et al (18 authors) 2022. Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing, Research Ideas and Outcomes (in press)

[19] Agosti et al. (submitted). Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Rio

GBIF 2021. Darwin Core Archives – How-to Guide, version 2.1, released on 8 February 2021, (contributions by Remsen D, Braak, K, Döring M, Robertson, T, Blissett M), Copenhagen: Global Biodiversity Information Facility, accessible online at: https://github.com/gbif/ipt/wiki/DwCAHowToGuide

Kalfatovic M 2010. BHL Australia Kick Off Meeting: Melbourne Museum. 1 June 2010. Melbourne, Australia. URL: https://www.slideshare.net/Kalfatovic/3-years-on-thebiodiversity-heritage-library?qid=3a0bdbbc-8b89-4260-a69d-93b58c8c6885&v=&b=&from_search=19

Meeus S, Addink W, Agosti D, Arvanitidis C, Balech B, Dillen M, Dimitrova M, González-Aranda JM, Holetschek J, Islam S, Jeppesen TS, Mietchen D, Nicolson N, Penev L, Robertson T, Ruch P, Trekels M, Groom Q 2022. Recommendations for interoperability among infrastructures. Research Ideas and Outcomes 8: e96180. doi: 10.3897/rio.8.e96180

Smirnova L, Mergen P, Groom Q, De Wever A, Penev L, Stoev P, Pe'er I, Runnel V, Camacho A, Vincent T, Agosti D, Arvanitidis C, Bonet F, Saarenmaa H 2016. Data sharing tools adopted by the European Biodiversity Observation Network Project. Research Ideas and Outcomes 2: e9390. doi: 10.3897/rio.2.e9390

Venkatesan A, Kim JH, Talo F, Ide-Smith M, Gobeill J, Carter J, Batista-Navarro R, Ananiadou S, Ruch P, McEntyre J. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. Wellcome Open Res. 2017 Jul 10;1:25. doi: 10.12688/wellcomeopenres.10210.2.

Appendix

Appendix 1.

Annotations available in TreatmentBank

The most common annotations

Annotation Type	Annotation Count	Document Count	Comment
emphasis	26349763	63580	emphasis in bold, italics, all-caps, small-caps, or combination thereof
taxonomicName	16298713	72727	taxonomic name
paragraph	15085096	73968	paragraph
td	11380660	37396	table cell
bibRefCitation	6761287	67775	citation of a bibliographic reference
author	6071803	71686	publisher in bibliographic reference
collectingCountry	3731595	57221	materials citation detail specifying the country a specimen was collected in
subSubSection	2980245	57194	subsection of a subsection or taxonomic treatment
bibRef	2748616	70739	bibliographic reference
year	2702350	71612	year of publication in bibliographic reference
title	2692258	71621	title in bibliographic reference
journalOrPublisher	2587220	71502	journal name or publisher in a bibliographic reference, when not possible to tell apart
collectionCode	2473382	41182	materials citation detail specifying the code of the institution or collection a specimen is deposited in
figureCitation	2433881	57098	in-text citation of a figure
tr	2350761	37707	table row
pagination	2202596	69366	pagination in a bibliographic reference
quantity	2072315	51859	generic quantity, with unit and conversion to associated SI unit

date	1997048	53486	date, with value normalized to ISO
part	1965892	54769	part designator in bibliographic reference, i.e., volume or issue number, or numero
normalizedToken	1831932	17849	token normalized to basic latin characters to simplify regex matching, bearing original value in attribute
collectingRegion	1750336	41491	materials citation detail specifying the region or first level administrative division a specimen was collected in
geoCoordinate	1630335	43195	geographic coordinate, mainly used as materials citation detail specifying the coordinates a specimen was collected at
specimenCount	1610892	40785	materials citation detail specifying the number of specimens collected together
typeStatus	1412222	51719	materials citation detail specifying the type status of a cited specimens
materialsCitation	1309437	38113	materials citation, container for detauls about cited specimen
collectorName	1209940	29952	materials citation detail specifying the person who collected a specimen
th	1158193	24287	table header cell
heading	1062183	55434	heading
location	1052142	35567	materials citation detail specifying the location a specimen was collected at
taxonomicNameLabel	992518	49109	status label going with taxonomic name in treatment taxon role, e.g. Marking new taxon or combination
collectingDate	878233	28724	materials citation detail specifying the date a specimen was collected
subSection	702015	72859	sub section of an article or book chapter
treatment	691914	55358	taxonomic treatment, special type of subSection
treatmentCitation	666104	17650	treatment citation
caption	648791	68710	caption of a figure or table
DOI	484444	32611	DOI in bibliographic reference
volume	462743	16202	volume number in bibliographic reference
collectingMunicipality	375862	30491	materials citation detail specifying the municipality a specimen was collected in
superScript	360557	27833	superscript
elevation	357969	20716	materials citation detail specifying the elevation a specimen was collected at
treatmentCitationGroup	346097	17275	group of treatment citations, sharing single taxon name
editor	344433	45606	editor in bibliographic reference

keyLead	330558	13877	lead in a taxonomic key
specimenCode	319806	12109	materials citation detail specifying the code a specimen is deposited under in a collection

All annotations

Annotation Type	Annotation Count	Document Count	Comment
abbreviation	551	59	an abbreviation that is associated with data
abbreviationData	740	97	the detail data associated with an abbreviation
abbreviationRange	222	67	range of abbreviations
abbreviationReference	299	5	reference to an abbreviation, implying details associated with abbreviation
accessDate	22834	9973	access date in bibliographic references to websites, etc.
accessionNumber	238826	16021	an accession number
author	6071803	71686	materials citation detail specifying the type status of a cited specimens
bedrock	1	1	materials citation detail specifying type of bedrock a specimen was found in, mainly for fossils
bibCitation	17696	449	in-line bibliographic citation
bibRef	2748616	70739	bibliographic reference
bibRefCitation	6761287	67775	citation of a bibliographic reference
bookContentInfo	211853	37587	number of pages, figures, plates, etc. In a bibliographic reference to a book
caption	648791	68710	caption of a figure or table
collectedFrom	13845	2129	materials citation detail specifying immediate environment a specimen was collected from
collectingCountry	3731595	57221	materials citation detail specifying the country a specimen was collected in

collectingCounty	175647	24072	materials citation detail specifying the county a specimen was collected in
collectingDate	878233	28724	materials citation detail specifying the date a specimen was collected
collectingMethod	64283	4889	materials citation detail specifying how a specimen was collected
collectingMunicipality	375862	30491	materials citation detail specifying the municipality a specimen was collected in
collectingPermit	9	7	materials citation detail specifying the permit a specimen was collected under
collectingRegion	1750336	41491	materials citation detail specifying the region or first level administrative division a specimen was collected in
collectionCode	2473382	41182	materials citation detail specifying the code of the institution or collection a specimen is deposited in
collectionCodeDefinition	12269	1741	definition of a collection code in materials and methods
collectionName	1	1	part of collection code definition
collectorName	1209940	29952	materials citation detail specifying the person who collected a specimen
crop	2	2	crop in plant usage
date	1997048	53486	date, with value normalized to ISO
determinerName	34624	2519	materials citation detail specifying the person who determined a specimen
docAuthor	96912	34157	document author in article head
docAuthorAffiliation	43797	13011	affiliation of document author in article head
docAuthorEmail	15202	6991	email address of document author in article head
docAuthorLSID	2074	674	LSID of document author in article head
docAuthorORCID	11344	4199	ORCID of document author in article head
docAuthorURL	22	15	personal URL of document author in article head
docEditor	19	9	document editor in article head
docIdDOI	2033	2031	document DOI in article head
docIdISSN	975	975	document ISSN in article head

docIdZooBank	1631	1631	ZooBank ID of document in article head
docIssue	12169	12169	issue number of document in article head
docJournal	4665	4665	parent journal of document in article head
docLocation	24	24	publisher location of document in article head
docNumero	8	8	numero of document in article head
docPagination	3661	3661	pagination of document in article head
docPubDate	4590	4590	exact publication date of document in article head
docPublicationUrl	447	447	URL of document in article head
docPublisher	14	14	publisher of document in article head
docRef	2284	2225	pagination of document in article head
docReference	1	1	full bibliographic reference to document given in article head
docSeriesInJournal	301	301	series within parent journal of document in article head
docTitle	53239	53071	title of document in article head
docVolume	3056	3056	volume number of document in article head
docVolumeTitle	11	11	title of parent volume (book or individually titled special issue of journal) in article head
docYear	7400	7400	year of publication of document in article head
DOI	484444	32611	DOI in bibliographic reference
editor	344433	45606	editor in bibliographic reference
elevation	357969	20716	materials citation detail specifying the elevation a specimen was collected at
emphasis	2634976 3	63580	emphasis in bold, italics, all-caps, small-caps, or combination thereof
figureCitation	2433881	57098	in-text citation of a figure
footnote	10439	4048	footnote
geneSequence	1426	388	gene sequence
geoCoordinate	1630335	43195	geographic coordinate, mainly used as materials citation detail specifying the coordinates a specimen was collected at
geologicalTimeScale	1207	45	materials citation detail specifying the geological time scale a specimen originated from, mainly for fossiles

heading	1062183	55434	heading
httpUri	145	25	the httpUri of cited materials if explicitly given in document text
insertion	30	14	explanatory insertion in text, most frequently in square brackets
issue	75521	5318	issue number in bibliographic reference
journal	10	5	journal name in a bibliographic reference
journalOrPublisher	2587220	71502	journal name or publisher in a bibliographic reference, when not possible to tell apart
key	11579	6132	taxonomic key
keyLead	330558	13877	lead in a taxonomic key
keyStep	169462	13863	step in taxonomic key
location	1052142	35567	materials citation detail specifying the location a specimen was collected at
locationDeviation	61506	6968	materials citation detail specifying the location a specimen was collected at relative to a more prominent landmark
materialsCitation	1309437	38113	materials citation, container for detauls about cited specimen
mods:affiliation	80387	21269	part of MODS metadata header in XML documents
mods:classification	72390	72389	part of MODS metadata header in XML documents
mods:date	74263	74263	part of MODS metadata header in XML documents
mods:dateIssued	640	640	part of MODS metadata header in XML documents
mods:dateOther	227	227	part of MODS metadata header in XML documents
mods:detail	165059	73713	part of MODS metadata header in XML documents
mods:end	74135	74134	part of MODS metadata header in XML documents
mods:extent	74136	74135	part of MODS metadata header in XML documents
mods:identifier	271969	74742	part of MODS metadata header in XML documents
mods:location	34256	34256	part of MODS metadata header in XML documents
mods:mods	74906	74906	part of MODS metadata header in XML documents
mods:name	226384	74906	part of MODS metadata header in XML documents
mods:nameIdentifier	46790	14680	part of MODS metadata header in XML documents

mods:namePart	226390	74906	part of MODS metadata header in XML documents
mods:number	163909	73713	part of MODS metadata header in XML documents
mods:originInfo	640	640	part of MODS metadata header in XML documents
mods:part	74735	74735	part of MODS metadata header in XML documents
mods:partNumber	1	1	part of MODS metadata header in XML documents
mods:place	511	511	part of MODS metadata header in XML documents
mods:placeTerm	511	511	part of MODS metadata header in XML documents
mods:publisher	638	638	part of MODS metadata header in XML documents
mods:relatedItem	74735	74735	part of MODS metadata header in XML documents
mods:role	226374	74906	part of MODS metadata header in XML documents
mods:roleTerm	226374	74906	part of MODS metadata header in XML documents
mods:start	74135	74135	part of MODS metadata header in XML documents
mods:subject	203	203	part of MODS metadata header in XML documents
mods:title	150794	74906	part of MODS metadata header in XML documents
mods:titleInfo	149644	74906	part of MODS metadata header in XML documents
mods:topic	10706	203	part of MODS metadata header in XML documents
mods:typeOfResource	74894	74894	part of MODS metadata header in XML documents
mods:url	34256	34256	part of MODS metadata header in XML documents
normalizedToken	1831932	17849	token normalized to basic latin characters to simplify regex matching, bearing original value in attribute
pageBreakToken	184009	16983	marker for token immediately following page break, indicating page boundaries in XML
pageNumber	582	169	page number in layout oriented XML
pageTitle	2450	403	page header or footer on layout oriented XML
pagination	2202596	69366	pagination in a bibliographic reference
paragraph	15085096	73968	paragraph

parenthesis	62	32	parenthesis, e.g. text box
part	1965892	54769	part designator in bibliographic reference, i.e., volume or issue number, or numero
plant	22	6	plant, in usage for crops, etc.
publicationUrl	103474	26355	publication URL in bibliographic reference
publisher	5	5	publisher in bibliographic reference
quantity	2072315	51859	generic quantity, with unit and conversion to associated SI unit
quote	312	27	part in quotes, mainly verbatim specimen labels
soilLayer	2	1	materials citation detail specifying the soil layer a specimen was collected from
specimenCode	319806	12109	materials citation detail specifying the code a specimen is deposited under in a collection
specimenCount	1610892	40785	materials citation detail specifying the number of specimens collected together
sub	2485	245	subscript, should have been transformed or removed
subScript	158873	8303	subscript
subSection	702015	72859	sub section of an article or book chapter
subSubSection	2980245	57194	sub section of a sub sestion or taxonomic treatment
sup	2718	441	superscript, should have been transformed or removed
superScript	360557	27833	superscript
table	118004	37741	table
tableCell	1	1	table cell in layout oriented XML
tableCitation	154886	26920	in-text citation of table
tableCol	16	3	table column in layout oriented XML
tableNote	19710	6732	note associated with table
tableRow	50	6	table row in layout oriented XML
taxonomicName	16298713	72727	taxonomic name
taxonomicNameLabel	992518	49109	status label going with taxonomic name in treatment taxon role, e.g. Marking new taxon or combination
taxonRecordFieldHeader	13	1	marker on table column headers specifying column semantics when transforming rows into record treatments

tbody	684	68	table body group of rows
td	11380660	37396	table cell
th	1158193	24287	table header cell
thead	877	438	table header group of rows
title	2692258	71621	title in bibliographic reference
tr	2350761	37707	table row
trait	1	1	morphological trait, name to be unified
traitTerm	408	6	morphological trait, name to be unified
treatment	691914	55358	taxonomic treatment, special type of subSection
treatmentCitation	666104	17650	treatment citation
treatmentCitationGroup	346097	17275	group of treatment citations, sharing single taxon name
treatmentCitationLabel	53	8	status label going with treatment citations, e.g. Marking new synonymies
treatmentDuplicate	24170	1842	duplicate treatment, renamed to take offline
typeStatus	1412222	51719	materials citation detail specifying the type status of a cited specimens
uri	15668	4921	generic URI
uuid	33540	5967	generic UUID, oftern used as precursor for URIs
verbatimSoil	4	2	materials citation detail containing the verbatim description of the soil a specimens was found in
verbatimVegetation	90	3	materials citation detail containing the verbatim description of the vegetation a specimens was found in
vernacularName	30472	173	vernacular name of a taxon
volume	462743	16202	volume number in bibliographic reference
volumeTitle	203383	47471	title of host volume in bibliographic reference, mainly book title for references to individual chapters
year	2702350	71612	year of publication in bibliographic reference