Web interface for ELIXIR Contextual Data ClearingHouse

Deliverable D8.3

28 January 2022

Authors:

Kessy Abarenkov[1], Allan Zirk[1], Guy Cochrane[2], Vishnukumar Kadhirvelu[2], Suran Jayathilaka[2], Joana Paupério[2], Olaf Banki[3], Jerry Lanfear[4,] Filipp Ivanov[1], Timo Piirmann[1], Raivo Pöhönen[1], Urmas Kõljalg[1]

Contributors:

1: University of Tartu Natural History Museum, Estonia
2: European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom
3: Sp2000
4: ELIXIR-Hub

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
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Web interface for ELIXIR Contextual Data ClearingHouse
Deliverable n°:	D8.3
Nature of the deliverable:	Other
Dissemination level:	Public
WP responsible:	WP8
Lead beneficiary:	University of Tartu
Citation:	Abarenkov, K., Zirk, A., Cochrane, G., Kadhirvelu, V., Jayathilaka, S., Paupério, J., Banki, O., Lanfear, J., Ivanov, F., Piirmann, T., Pöhönen, R. & Kõljalg, U. (2022). *Web interface for ELIXIR Contextual Data ClearingHouse*. Deliverable D8.3 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 9
Actual submission date:	28 January 2022

Deliverable status:

Version	**Status**	**Date**	**Author(s)**
1.0	Draft	19 January 2022	1: University of Tartu Natural History Museum, Estonia 2: European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom 3: Sp2000 4: ELIXIR-Hub
2.0	Review	26 January 2022	Lyubomir Penev Quentin Groom
3.0	Submission	28 January 2022	University of Tartu Natural History Museum, Estonia

Table of contents

Preface

Third-party annotations are a valuable resource to improve the quality of public DNA sequences. For example, sequences in the International Nucleotide Sequence Databases Collaboration (INSDC)[1] often lack important features like species level identification, information associated with habitat, locality, country, coordinates, interactions between taxa etc. Third-party annotations have their own specific challenges. For example, annotations can be inaccurate and therefore must be open for permanent data management. Further, every DNA sequence (except sequences from type material) can carry different species names which must be recorded as equal scientific hypotheses. The PlutoF[2] platform provides such data management services for third-party annotations.

ELIXIR Contextual Data ClearingHouse[3] offers a lightweight and simple RESTful API to enable extension, correction and improvement of publicly available annotations on sample and sequence records available in ELIXIR data resources.

The work of linking these two components – web interface provided by the PlutoF platform and the ELIXIR Contextual Data ClearingHouse APIs – to allow user-friendly and effortless reporting of errors and gaps in sequenced material source annotations, has been carried out as part of the BiCIKL Project[4] and is described in this document, the Deliverable D8.3 of BiCIKL: Web interface for ELIXIR Contextual Data ClearingHouse.

Summary

This deliverable report includes description of the work steps towards building a web interface for the reporting of errors and gaps in sequenced material source annotations as part of the Task 8.3 of BiCIKL. Beta version of the web interface has been published and is available for the registered users of PlutoF platform.

List of abbreviations

ENA	European Nucleotide Archive
INSDC	International Nucleotide Sequence Databases Collaboration

[1] https://www.insdc.org/
[2] https://plutof.ut.ee
[3] https://www.ebi.ac.uk/ena/clearinghouse/api/
[4] https://bicikl-project.eu/

1. Background, scientific objectives, and state of the art

1.1. ENA

The European Nucleotide Archive (ENA) is an open platform for the management, sharing, integration and dissemination of sequence data. The ENA is the European node of the INSDC and offers extensive public domain data for over 1.5 million species. ENA comprises a comprehensive databasing infrastructure for the archiving of petabytes of sequence data and associated metadata, and a portfolio of tools and services for the management of sequence data. These tools include the Webin data submission and validation application, that is widely used, and sophisticated data discovery and retrieval tools.
ENA holds a large amount of metadata relating to the sample source for an organism, as a culture collection or a natural history collection. However, for a number of records these annotations may be incomplete (sequences not linked to their sources), ambiguous (may lead to multiple endpoints) or even inaccurate. Therefore, there is the need to facilitate data update cycles when these omissions and inaccuracies are detected.
The ELIXIR Contextual Data Clearinghouse (subsequently referred to as the Clearinghouse) allows those with corrections and additions to current metadata, such as information on material sources or sequencing library details, to feed this information to primary repositories. The data repositories, such as ENA, can then access the annotations made, review them and display the updates if appropriate.

1.2. PlutoF

PlutoF is an online data management platform and computing service provider for biology and related disciplines. Registered users can enter and manage a wide range of data, e.g. taxon occurrences, metabarcoding data, taxon classifications, traits, lab data, etc. It also features an annotation module where third-party annotations (on material source, geolocation and habitat, taxonomic identifications, interacting taxa, etc.) can be added to any collection specimen, living culture or DNA sequence record.

2. Implementation of the workflow in PlutoF

The work to be conducted under Task 8.3 of BiCIKL was divided into multiple stages –

- Stage 1: Implement third-party annotation submission workflow from PlutoF workbench to the Clearinghouse
- Stage 2: Implement the user-initiated process of fetching International Nucleotide Sequence Database (INSD) sequence data to be incorporated into PlutoF for annotating purposes
- Stage 3: Provide public services for retrieving Clearinghouse third-party annotations through community specific portals

Stage 1 focused on implementing third-party annotation submission workflow from PlutoF workbench to the Clearinghouse. While PlutoF workbench currently provides third-party annotation options for all its molecular sequence data (regularly updated dataset of ribosomal DNA Internal Transcribed Spacer, Small Subunit, and Large Subunit sequences downloaded from INSD), our specific purpose in Stage 1 was to make all INSD sequence annotations added through the PlutoF workbench available through the Clearinghouse.

2.1. General data flow

International Nucleotide Sequence Database data and metadata are downloaded from INSD using NCBI's E-utilities[5] on a regular basis (Image 1). These data, together with PlutoF own database records prior to submitting to INSDC, are stored and made available for third-party annotating in PlutoF.

Annotation workflow steps:

a) User annotates sequence metadata by clicking on the "Annotate" link in the sequence view.
b) An Annotation Proposal will be created, and verification notification sent out to the designated reviewer.
c) The reviewer either accepts the Annotation Proposal or rejects it with a comment.
d) If Annotation Proposal is accepted, annotated fields that could be mapped to ENA fields are pushed to the Clearinghouse using their RESTful API.

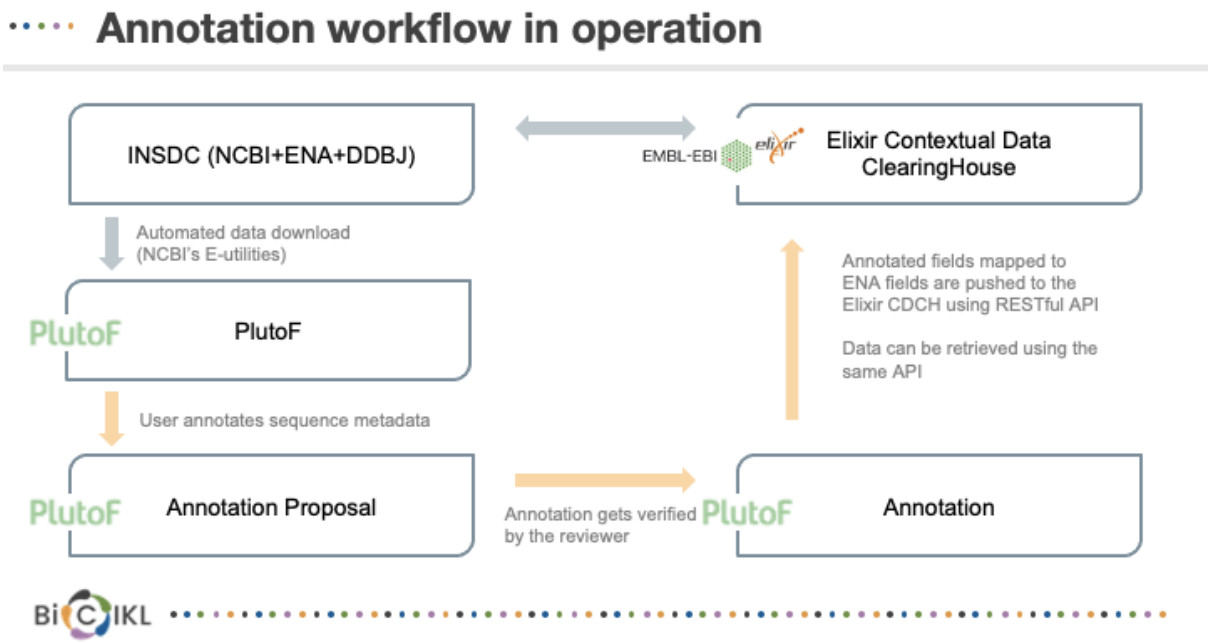

Image 1. Graph describing how annotations are added and verified in PlutoF and sent to the ELIXIR Clearinghouse.

[5] https://eutils.ncbi.nlm.nih.gov/

2.2. Mapping of the PlutoF and ENA fields available for annotating

We identified 281 fields that already existed in PlutoF for collecting metadata about DNA sequence and its source. Out of the total 281 PlutoF fields we were able to map 32 to their corresponding ENA fields (total count of the corresponding ENA fields: 16, Table 1) using the ENA Features and Qualifiers table[6] for reference.

Table 1: Mapping table between the PlutoF and ENA fields open for third-party annotation.

No.	ENA feature	ENA qualifier	PlutoF module	PlutoF field	Example
1	source	db_xref	Sequence	Sequence ID	/db_xref="UNITE:UDB000157"
2	source	isolation_source	Sequence	Isolation source	/isolation_source="plant leaf"
3	source	PCR_primers	Sequence	Forward primer name	/PCR_primers="fwd_name: ITS1F, fwd_seq: CTTGGTCATTTAGAGGAAGTAA, rev_name: ITS4B, rev_seq: CAGGAGACTTGTACACGGTCCAG"
4	source	PCR_primers	Sequence	Forward primer sequence	/PCR_primers="fwd_seq: CTTGGTCATTTAGAGGAAGTAA"
5	source	PCR_primers	Sequence	Reverse primer name	/PCR_primers="rev_name: ITS4B, rev_seq: CAGGAGACTTGTACACGGTCCAG"
6	source	PCR_primers	Sequence	Reverse primer sequence	/PCR_primers="rev_seq: CAGGAGACTTGTACACGGTCCAG"
7	source	note	Sequence	Chimeric	/note="This sequence is chimeric"
8	source	note	Sequence	Low quality	/note="This sequence is of low quality"
9	source	collection_date	Sequence	Sampling event.Timespan begin	/collection_date="2021-09-28"
10	source	collection_date	Sequence	Sampling event.Timespan end	/collection_date="2021-09-28/2021-09-29"
11	source	collected_by	Sequence	Sampling event.Collected by	/collected_by="Leho Tedersoo"
12	source	lat_lon	Sequence	Sampling event.Sampling area.Latitude	/lat_lon="47.94 N 28.12 W"
13	source	lat_lon	Sequence	Sampling event.Sampling area.Longitude	/lat_lon="47.94 N 28.12 W"
14	source	country	Sequence	Sampling event.Sampling area.Country	/country="Canada"

[6] https://www.ebi.ac.uk/ena/WebFeat/

15	source	country	Sequence	Sampling event.Sampling area.State	/country="Canada:Vancouver"
16	source	country	Sequence	Sampling event.Sampling area.District	/country="Estonia:Harju district"
17	source	country	Sequence	Sampling event.Sampling area.Commune or City	/country="Estonia:Harju district, Tallinn"
18	source	country	Sequence	Sampling event.Sampling area.Locality text	/country="Estonia:Harju district, Tallinn, near the harbour"
19	source	altitude	Sequence	Sampling event.Sampling area.Elevation min.Value	/altitude="320.14 m"
20	source	altitude	Sequence	Sampling event.Sampling area.Elevation max.Value	/altitude="180 m/250 m"
21	source	altitude	Sequence	Sampling event.Sampling area.Depth min.Value	/altitude="-100 m"
22	source	altitude	Sequence	Sampling event.Sampling area.Depth max.Value	/altitude="-100 m/-50 m"
23	source	organism; db_xref	Sequence	Determination.Taxon name	/organism="Boletus edulis"; /db_xref="taxon:36056"
24	source	type_material	Sequence	Determination.Typification	/type_material="holotype of Boletus edulis"
25	source	identified_by	Sequence	Determination.Identified by	/identified_by="Urmas Kõljalg"
26	source	bio_material	MaterialSample	Material sample ID	/bio_material=TUE001234
27	source	host	MaterialSample	Interaction.Taxon	/host="Alnus sp"
28	source	host	MaterialSample	Interaction.Interacting taxon type	/host="Alnus sp"
29	source	specimen_voucher	Specimen	Specimen ID	/specimen_voucher=TU<EST>:TUF001234
30	source	specimen_voucher	Specimen	Subcode	/specimen_voucher=TU<EST>:TUF001234.1
31	source	culture_collection	LivingSpecimen	Code	/culture_collection=TFC001234
32	source	culture_collection	LivingSpecimen	Subcode	/culture_collection=TFC001234.1

2.3. Setting up and verifying PlutoF annotation workflow

We then selected a set of third-party annotation use-cases to set up and verify the annotation workflow in PlutoF from the user's perspective. The list of use cases checked together with their status and comments is available in Table 2.

Table 2: List of use cases to set up and verify the PlutoF annotation workflow.

Use case description	Status	Comment
Annotate project fields (edit existing project metadata, mark study as published)	Cannot be done	PlutoF project metadata are not available for third-party annotation. If user wants to link DNA sequences under one project to published literature reference, it should be done by linking project/sequences to Reference object through the Associated Data panel.
Annotate project fields (move under another existing INSD project)	Cannot be done	No need to move INSD sequences between different INSD projects which are used for grouping data according to INSD submission.
Annotate project fields (move under new user-created project)	Cannot be done	No need to move INSD sequences under user's own project.
Annotate locality fields (e.g., specify country name)	Can be done	
Annotate sampling event fields (e.g., specify collecting date)	Can be done	
Annotate sequence traits	Can be done	No fields among the current PlutoF sequence traits that could be mapped to ENA fields.
Annotate data linked to sequence model (e.g., set quality/chimeric status, add sequenced regions)	Can be done	
Link associated data (add link to reference, add external link, add keywords)	Can be done	No fields among the current PlutoF Associated Data form that could be mapped to ENA fields.
Annotate source (to existing specimen/culture/materialsample)	Can be done	Can be done via public linking to existing Source object.
Annotate source (to user created specimen/culture/materialsample)	Can be done	Can be done via public linking to new Source object.
Add taxon identification	Can be done	
Add taxon interactions	Can be done	

2.4. Specific annotation use cases

Specific annotation use-cases are covered in the PlutoF third-party annotations user manual[7]. Since the development of the described online web interface is to be continued throughout the BiCIKL project, the PlutoF user manual will be the document where all updates to the interface will be published.

The PlutoF annotation module allows annotation of the following sequence metadata fields (grouped into wider categories):

References

In many occasions reference information for INSD sequence record indicates that the study this sequence originates is unpublished. Often studies get published after sequence

[7] https://plutof.ut.ee/assets/varia/manuals/docs/annotations_manual_en.pdf

submission to INSD but the status of the study remains unchanged by the authors. It is possible to indicate that a specific sequence is linked to a published study by linking this sequence with the PlutoF Reference object. Steps to add this information:

a) Search for existing reference object in PlutoF Reference search module[8]
b) If reference was not found, add new reference using Reference Add form[9]. Journal articles can be either imported using DOI or inserted manually.
c) Use this reference when annotating DNA sequences (*Associated Data => References*).
d) Submit annotation by clicking "Annotate". Annotations to associated references are currently not sent to the Clearinghouse but are stored and made available to PlutoF users, therefore clicking "Annotate to ENA" is not needed here.

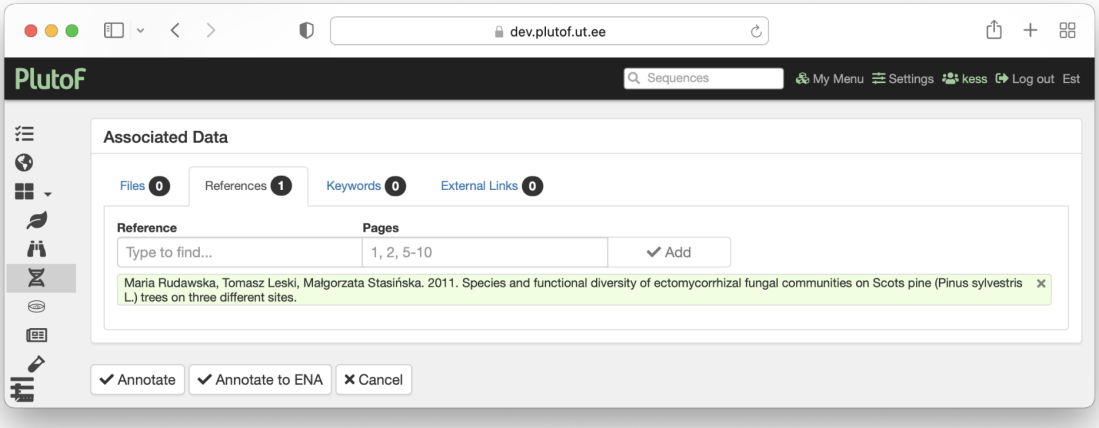

Image 2. Example form for adding and linking up-to-date Reference information to INSD sequence FJ158075.

Locality fields

Annotate Locality data (/lat_lon, /country) while in Sequence Annotate view "Area and Event" panel.

[8] https://plutof.ut.ee/#/search?module=reference
[9] https://plutof.ut.ee/#/reference/add

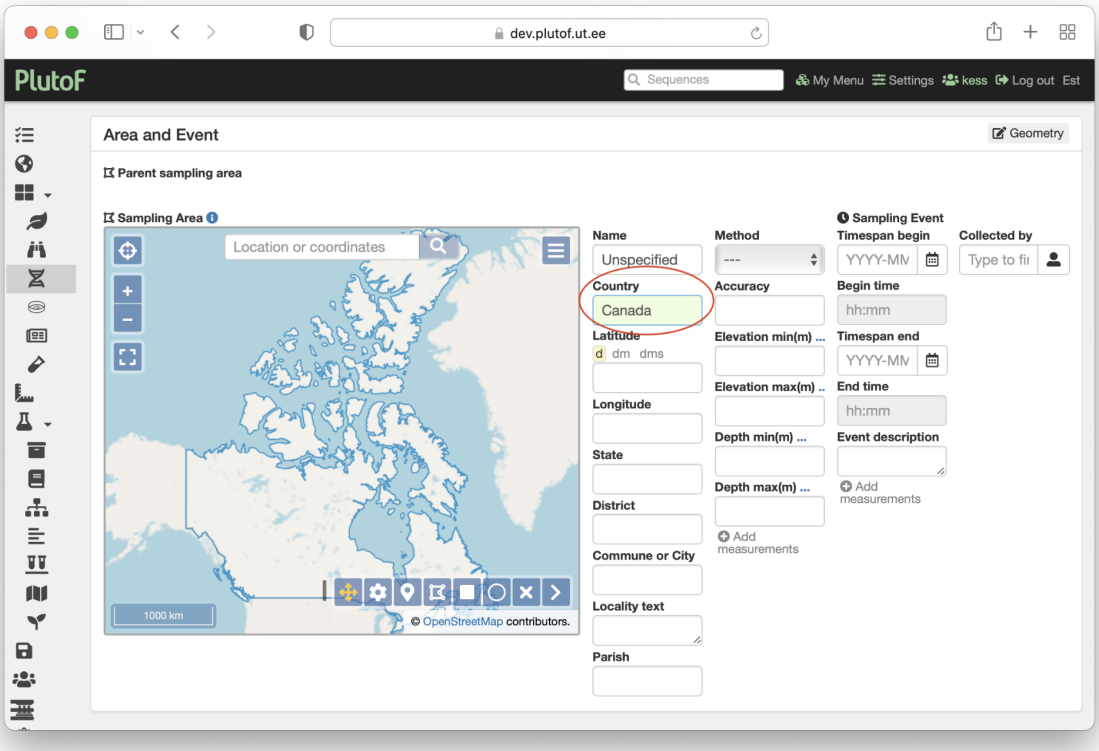

Image 3. Example form for adding up-to-date Locality information (by changing country name from Unspecified to Canada) to INSD sequence MH118168.

Sampling event fields

Annotate Event data (/collection_date, /collected_by, /altitude) while in the Sequence Annotate view “Area and Event” panel.

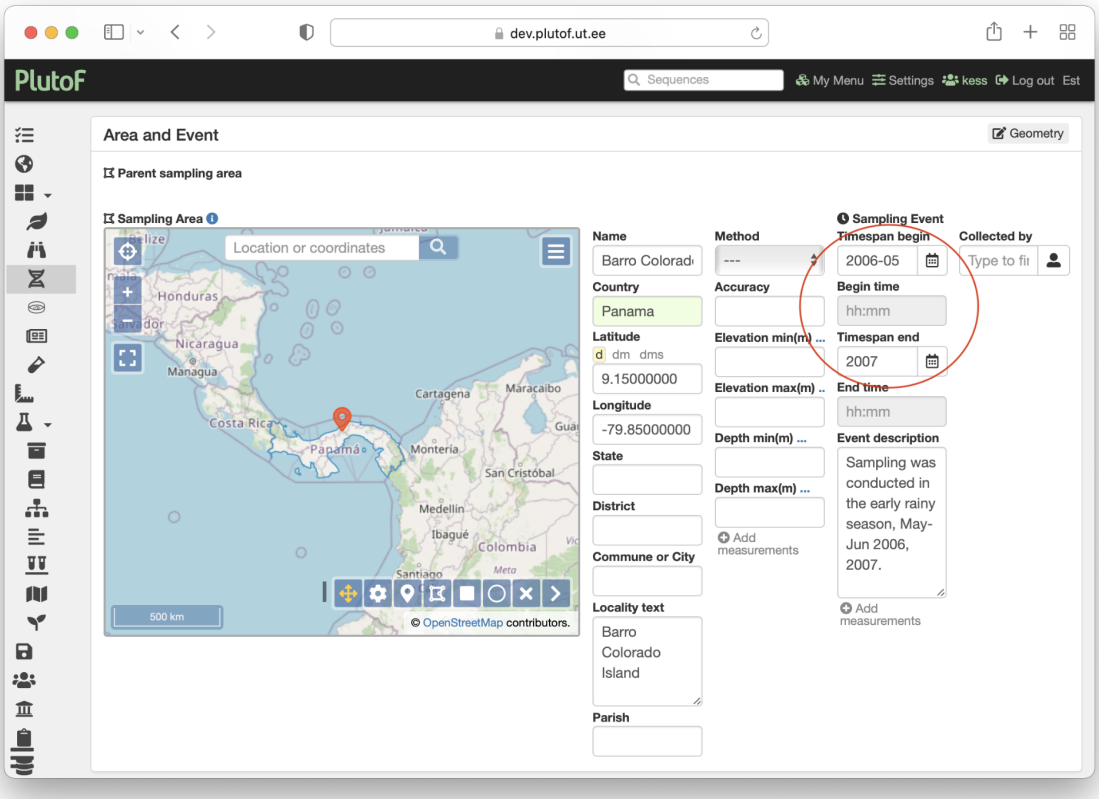

Image 4. Example form for adding up-to-date Sampling event information (by specifying collection date) to INSD sequence EU686795.

Fields directly linked to sequence

Annotate sequence metadata (/isolation_source, /PCR_primers, /note) while in the Sequence Annotate view “General Data” panel.

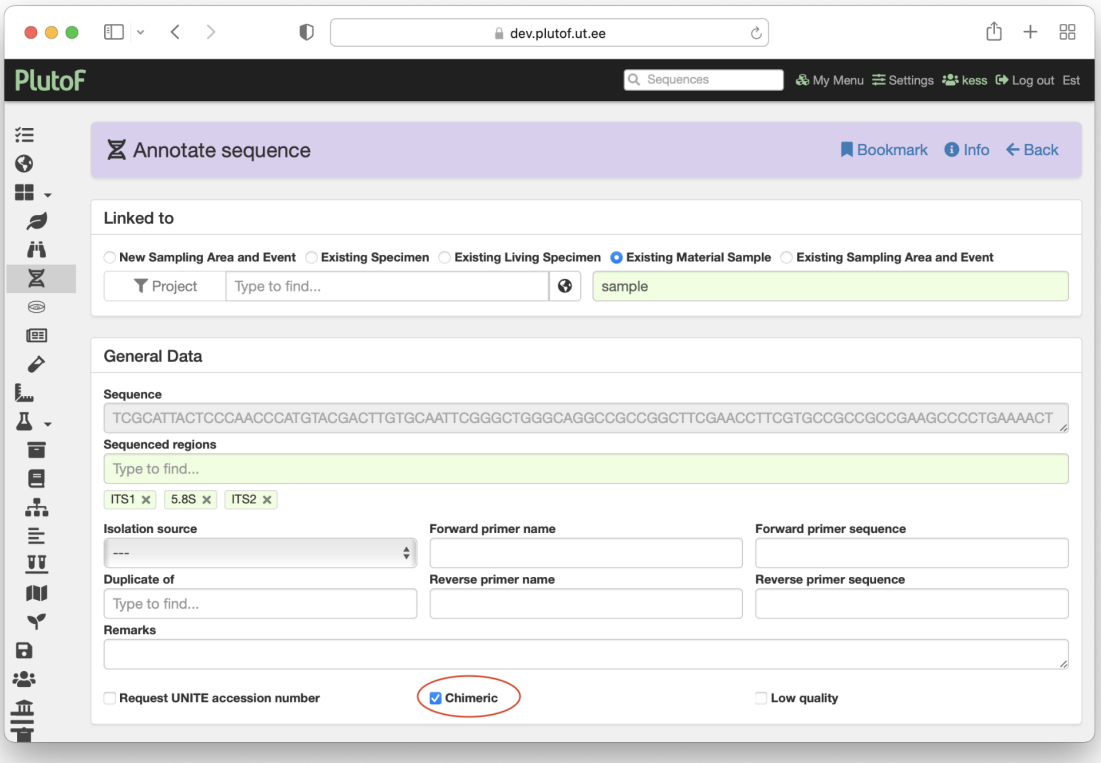

Image 5. Example form for flagging sequence INSD sequence FJ884118 as chimeric.

Linked data

Add linked data (references, external links, keywords) while in the Sequence Annotate view “Associated Data” panel. These annotations will not be sent to the Clearinghouse but will be stored and made available to PlutoF users.

Source (link to voucher specimen, culture or material sample)

Annotate sequence Source (/specimen_voucher, /bio_material, /culture_collection) while in the Sequence Annotate view “Linked to” panel.

It is possible to link sequences with new Source objects (such as specimens, cultures or material samples) which can be added and stored in PlutoF as individual Source records.

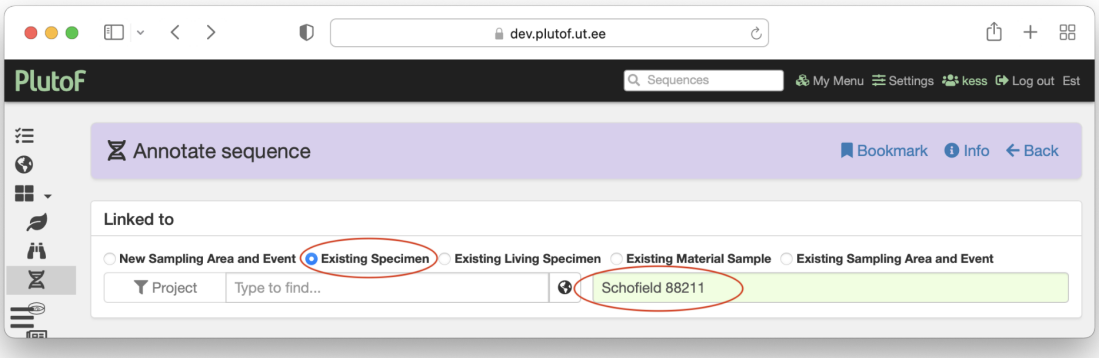

Image 6. Example for linking new Source record to INSD sequence JF734610.

When changing the source from "Existing Sampling Area and Event" (original data downloaded from INSD) to "Existing Specimen/Living Specimen/Material Sample", you will be prompted with the question if you want to use the source's sampling event or create a new one. If you (1) do not need OR (2) need and have access to editing the Source record, click on "Use Source's Event.

Taxon identifications

Reidentifications (/organism, /db_xref) to INSD sequences can be added by creating a new identification record while in the sequence view "Identifications=>Edit" panel.

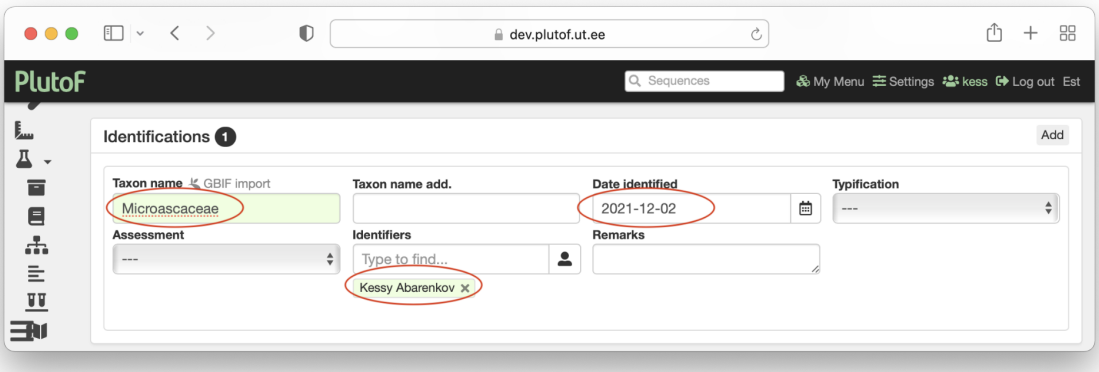

Image 7. Example form for adding new INSD sequence identification for FJ524321.

2.5. Submitting the annotations

Third-party annotations will be pushed to ENA by clicking the "Annotate to ENA" button. The user will be prompted with the annotation summary and additional metadata fields (e.g. Assertion Evidence and Comment; see Image 8) requested by the Clearinghouse API for those annotated fields that could be mapped to ENA fields (Table 1 in Section 2.2 of this document).

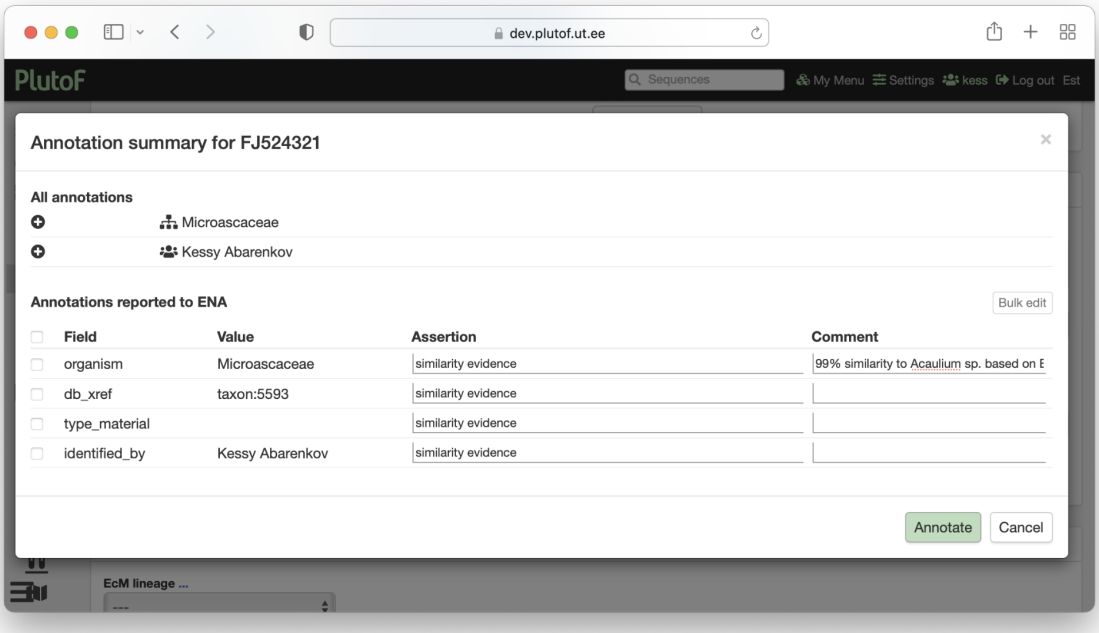

Image 8. Example view of the annotation summary page when submitting a new identification to ENA (example record: FJ524321).

Users' annotations can be found on the sequence record page inside the Annotation Proposals panel while changes added during the annotation are shown inside the History panel (Image 9).

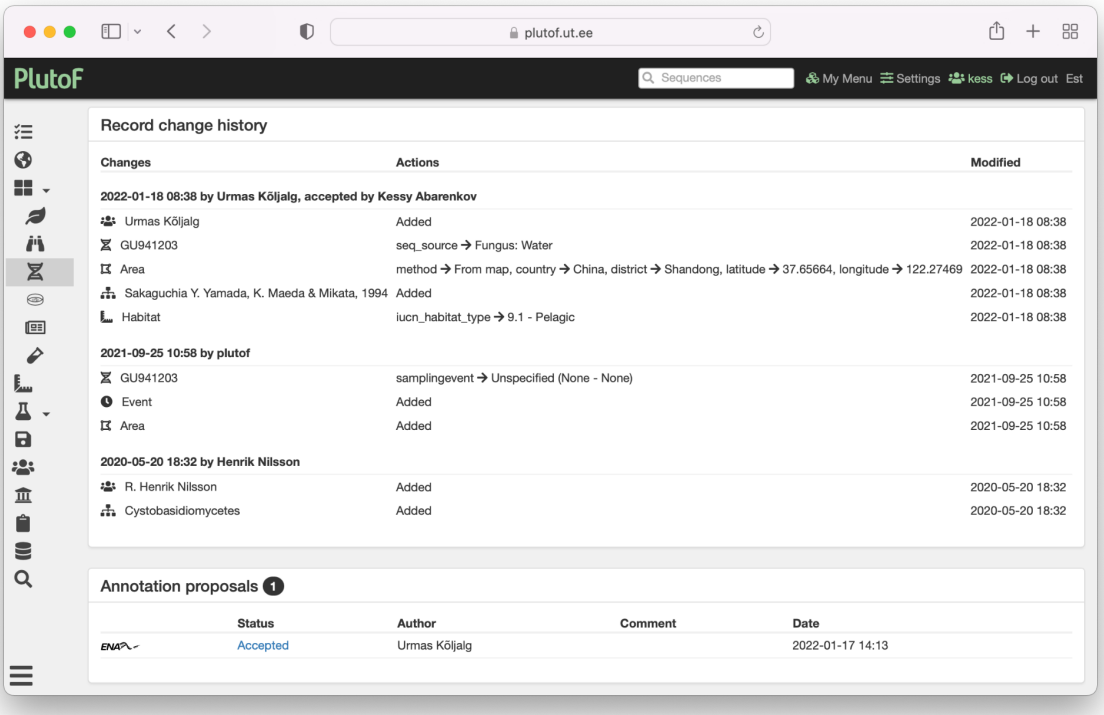

Image 9. Example view of the sequence record page with Annotations Proposals and History panel shown (example record: GU941203).

3. Take-up

Since September, 2021 when Task 8.3 started, 3 054 third-party annotation submissions by 16 users have been carried out in PlutoF using the annotation module (data from 18.01.2022). Out of these, 48 annotated attributes from 9 records[10] have been reported to the Clearinghouse starting from 17.01.2022 when the functionality to push annotations to Clearinghouse instead of storing in PlutoF was fully enabled.

The results of the work at Task 8.3 in BiCIKL was presented at the TDWG conference (Abarenkov et al. 2021).

4. Next steps

- March 2022: Implementation of the user-initiated process of fetching INSD sequence data to be incorporated into PlutoF for annotation purposes (Stage 2 in Section 2 of this document).
- March 2022: Implementation of new trait ontologies (biological interactions and nutritional modes) needed by the annotating communities.
- March 2022: Updated version of the users manual released.
- March 2022: Implementation of the process of revision and approval of third-party annotations into ENA records (Stage 3 in Section 2 of this document).
- May 2022: Updated version of the online annotation interface released.

5. Acknowledgements

We want to thank all the partners involved in Task 8.3 in BiCIKL for their input – discussions and recommendations for finding solutions to several occurring issues with the implementation of third-party annotation workflow.

6. References

Abarenkov K, Zirk A, Põldmaa K, Piirmann T, Pöhönen R, Ivanov F, Adojaan K, Kõljalg U (2021) Third-party Annotations: Linking PlutoF platform and the ELIXIR Contextual Data ClearingHouse for the reporting of source material annotation gaps and inaccuracies. Biodiversity Information Science and Standards 5: e74249. https://doi.org/10.3897/biss.5.74249

[10] https://www.ebi.ac.uk/ena/clearinghouse/api/curations?providerName=plutof&limit=100