PlutoF

Third-party annotations

User manual

(Last updated: 18.01.2022)

Table of contents

1. Introduction

Third-party annotations are a valuable resource to improve the quality of public DNA sequences. For example, sequences in International Nucleotide Sequence Databases Collaboration (INSDC) often lack important features like taxon interactions, species level identification, information associated with habitat, locality, country, coordinates, 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 databased as equal scientific hypotheses. PlutoF platform provides such data management services for third-party annotations.

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. PlutoF annotations are linked to sequence and sample records stored in INSD databases through operating the ELIXIR Contextual Data ClearingHouse (CDCH). CDCH offers a light and simple RESTful API to enable extension, correction and improvement of publicly available annotations on sample and sequence records available in ELIXIR data resources.

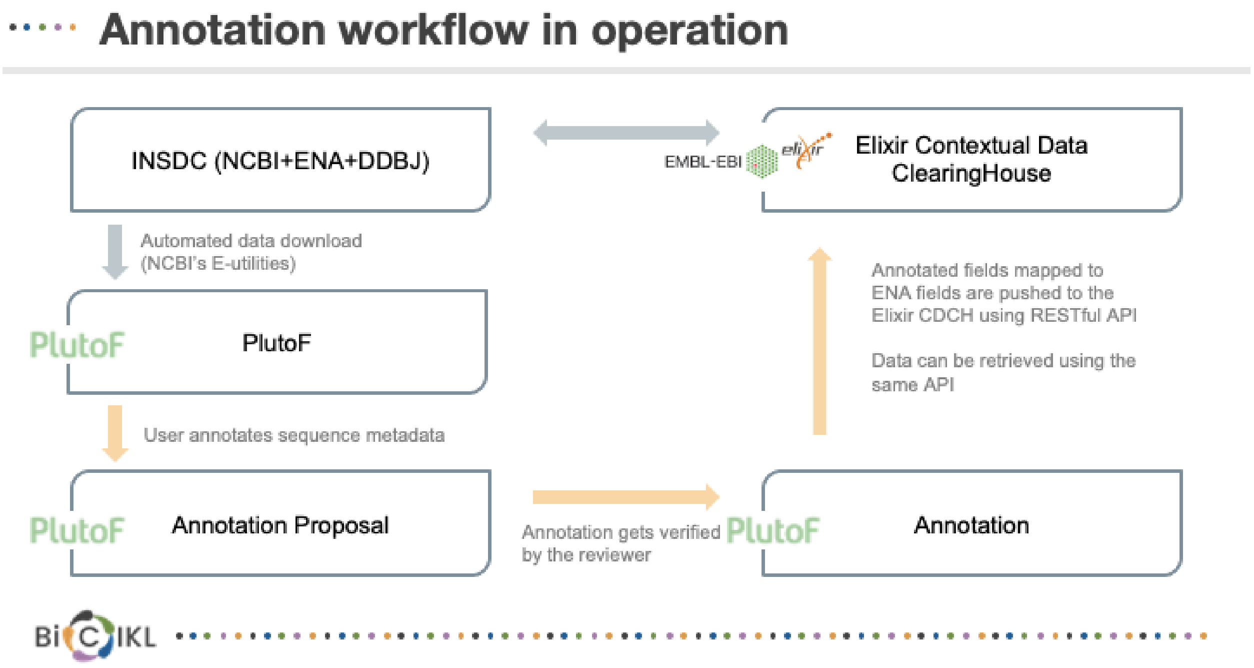
The work of linking these two components - web interface provided by the PlutoF platform and CDCH APIs – to allow user-friendly and effortless reporting of errors and gaps in sequenced material source annotations, has been carried out as part of the BiCIKL Project (https://bicikl-project.eu/).

2. General data flow

INSD sequence data and metadata are downloaded from INSD using NCBI's E-utilities on a regular basis (Image 1). These data 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) Annotation Proposal will be created, and verification notification sent out to the designated reviewer.
c) 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 INSD fields are pushed to the Elixir CDCH using their RESTful API (https://www.ebi.ac.uk/ena/clearinghouse/api/).



Annotation workflow in operation

INSDC (NCBI+ENA+DDBJ)

EMBL-EBI

elixir

Elixir Contextual Data ClearingHouse

Automated data download (NCBI's E-utilities)

PlutoF

PlutoF

Annotated fields mapped to ENA fields are pushed to the Elixir CDCH using RESTful API

Data can be retrieved using the same API

User annotates sequence metadata

PlutoF

Annotation Proposal

Annotation gets verified by the reviewer

PlutoF

Annotation

BiCIKL

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

3. Specific annotation use cases

PlutoF annotation module allows to annotate the following sequence metadata fields (grouped into wider categories) -

3.1. Reference(s)

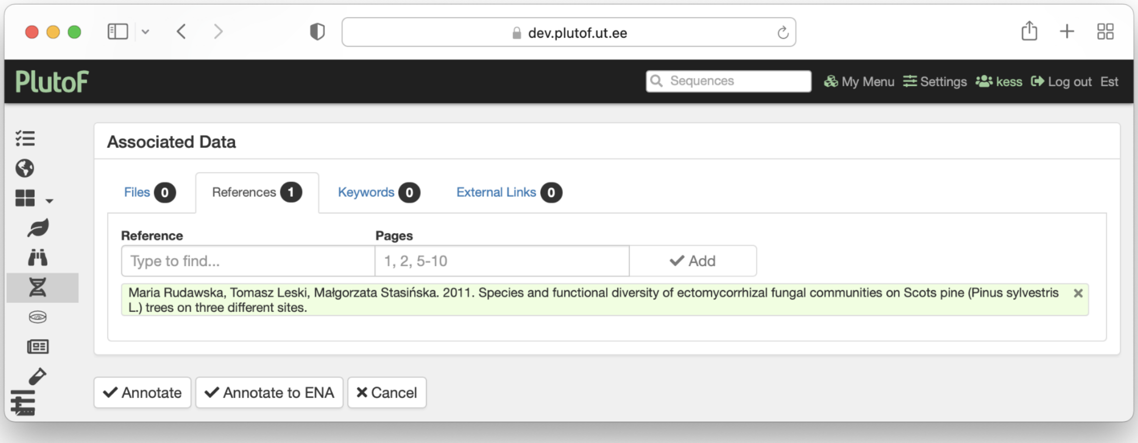
In many occasions Reference information for INSD sequence record indicates that the study this sequence originates is unpublished. Often studies get published after sequence submission to INSD. It is possible to indicate that a specific sequence is linked to a published study by linking this sequence with the PlutoF Reference object. Steps to add this information -

a) Search for existing reference object in PlutoF Reference search module (https://plutof.ut.ee/#/search?module=reference)

b) If reference was not found, add new reference using Reference Add form (https://plutof.ut.ee/#/reference/add). 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 will not be sent to Elixir CDCH but will be stored and made available to PlutoF users, therefore clicking “Annotate to ENA” is not needed here.



Associated Data

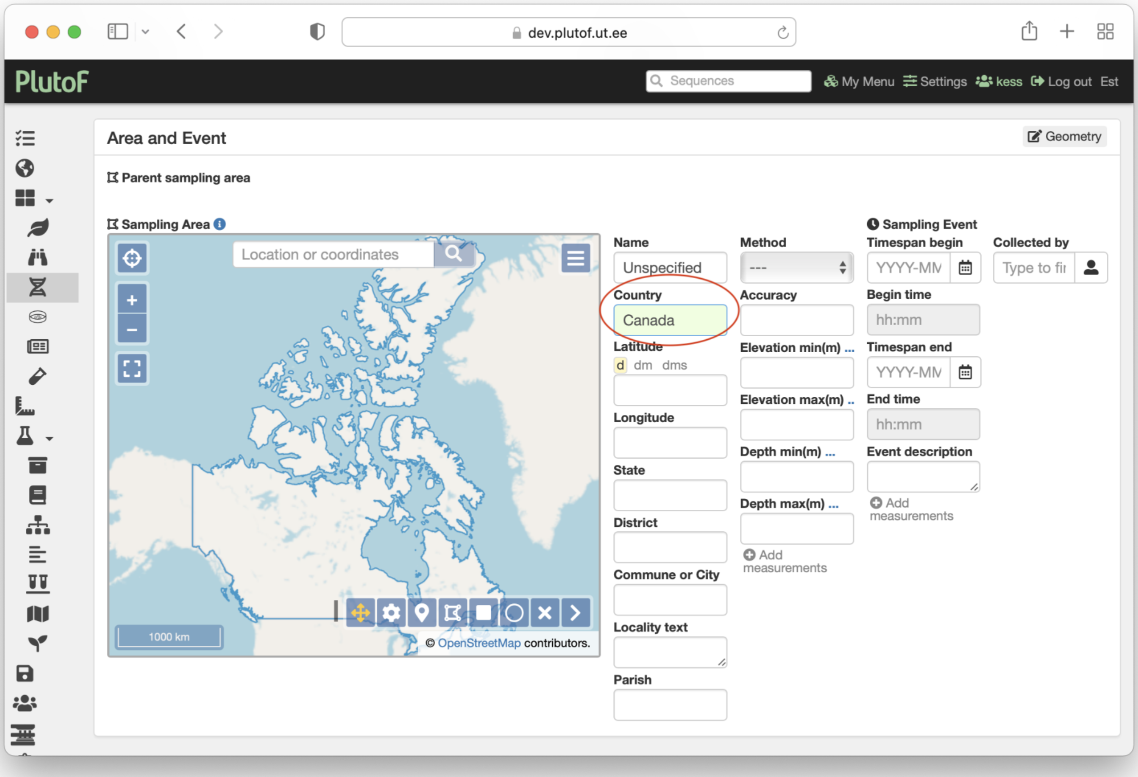
Reference	Pages
Type to find...	1, 2, 5-10

Maria Rudawska, Tomasz Leski, Małgorzata Stasińska. 2011. Species and functional diversity of ectomycorrhizal fungal communities on Scots pine (Pinus sylvestris L.) trees on three different sites.

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

3.2. Locality fields

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



Area and Event

Name: Unspecified

Country: Canada

Latitude: d dm dms

Longitude

State

District

Commune or City

Locality text

Parish

Method: ---

Accuracy

Elevation min(m) ...

Elevation max(m) ..

Depth min(m) ...

Depth max(m) ...

Timespan begin: YYYY-MM

Begin time: hh:mm

Timespan end: YYYY-MM

End time: hh:mm

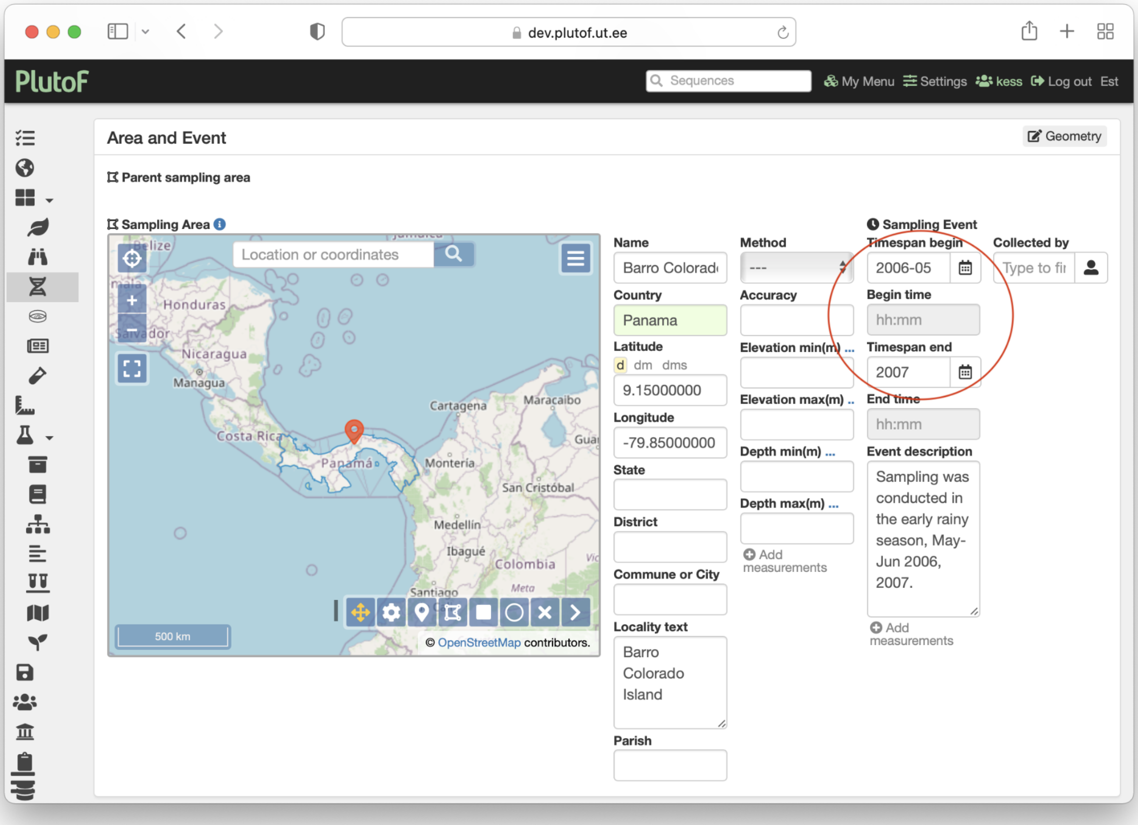
Event description

Collected by: Type to fi

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

3.3. Sampling event fields

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



Area and Event

Parent sampling area

Sampling Area

Sampling Event

Name	Barro Colorad
Country	Panama
Latitude (d dm dms)	9.15000000
Longitude	-79.85000000
State	
District	
Commune or City	
Locality text	Barro Colorado Island
Parish	
Method	---
Accuracy	
Elevation min(m) ...	
Elevation max(m) ..	
Depth min(m) ...	
Depth max(m) ...	
Timespan begin	2006-05
Begin time	hh:mm
Timespan end	2007
End time	hh:mm
Event description	Sampling was conducted in the early rainy season, May-Jun 2006, 2007.
Collected by	Type to fi

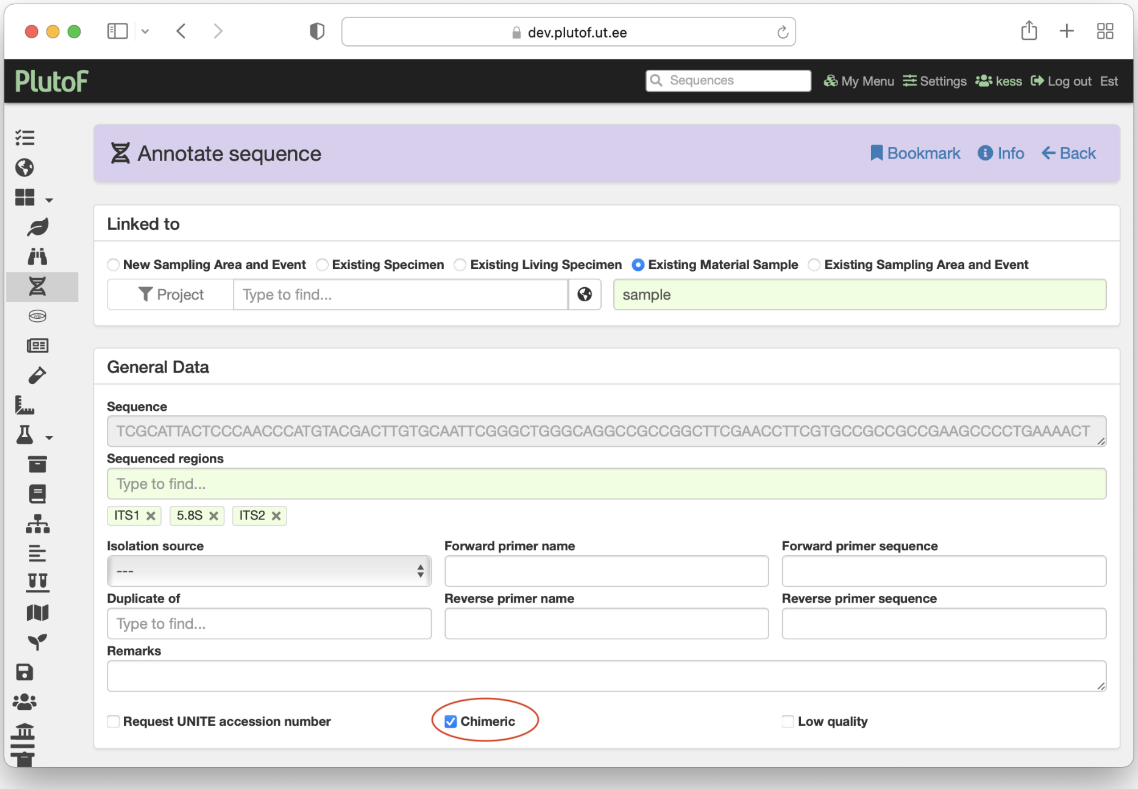
Add measurements

Add measurements

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

3.4. Fields directly linked to sequence model

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



Annotate sequence

Linked to

○ New Sampling Area and Event ○ Existing Specimen ○ Existing Living Specimen ● Existing Material Sample ○ Existing Sampling Area and Event

Project: Type to find...

sample

General Data

Sequence: TCGCATTACTCCCAACCCATGTACGACTTGTGCAATTCGGGCTGGGCAGGCCGCCGGCTTCGAACCTTCGTGCCGCCGCCGAAGCCCCTGAAAACT

Sequenced regions: Type to find...

ITS1 ✕ 5.8S ✕ ITS2 ✕

Isolation source	Forward primer name	Forward primer sequence

Duplicate of	Reverse primer name	Reverse primer sequence
Type to find...		

Remarks:

□ Request UNITE accession number ☑ Chimeric □ Low quality

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

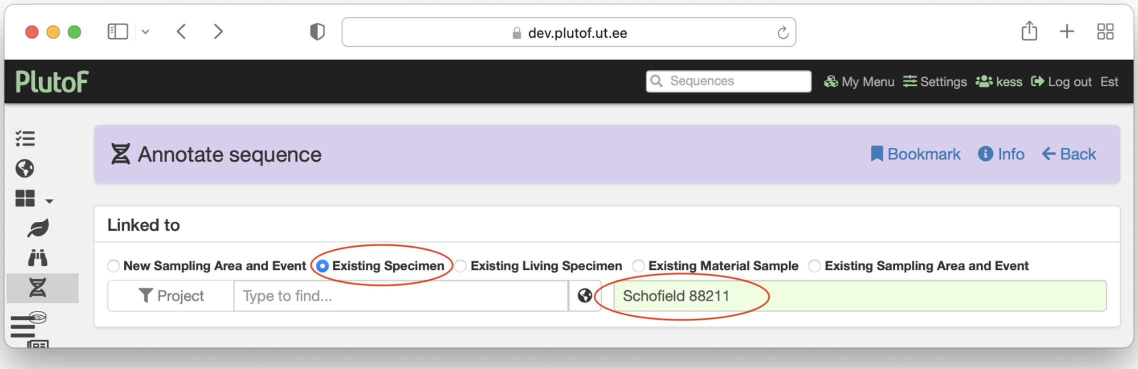
3.5. Linked data

Add linked data (references, external links, keywords) while in the Sequence Annotate view “Associated Data” panel (see p3.1. for a more detailed example). These annotations will not be sent to Elixir CDCH but will be stored and made available to PlutoF users.

3.6. 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.



Annotate sequence

Linked to

○ New Sampling Area and Event ● Existing Specimen ○ Existing Living Specimen ○ Existing Material Sample ○ Existing Sampling Area and Event

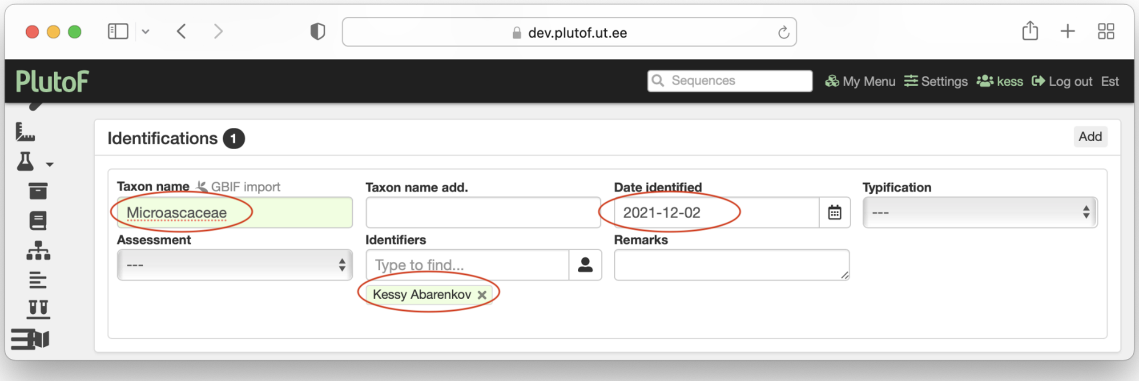
Project: Type to find... | Schofield 88211

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. In case you 1) do not need OR 2) need and have access to editing the Source record, click on “Use Source’s Event).

3.7. 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.



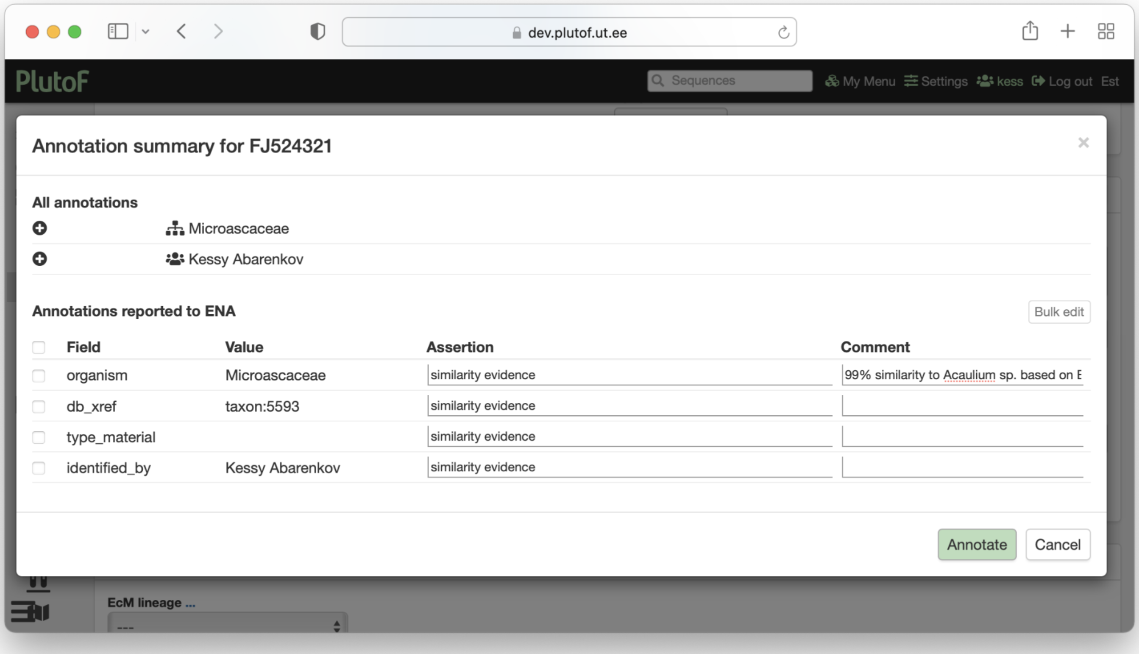
Identifications 1

Taxon name: Microascaceae	Taxon name add.:	Date identified: 2021-12-02	Typification: ---
Assessment: ---	Identifiers: Type to find... Kessy Abarenkov ×	Remarks:	

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

4. “Annotate to ENA” submission

Third-party annotations can be sent to ENA by clicking the “Annotate to ENA” button. User will be prompted with the Annotation summary and additional metadata fields (e.g. Assertion Evidence and Comment; see Image 8) requested by the Elixir CDCH API for those annotated fields that could be mapped to ENA fields (Supplementary table 1).



Annotation summary for FJ524321

All annotations

Microascaceae

Kessy Abarenkov

Annotations reported to ENA

	Field	Value	Assertion	Comment
□	organism	Microascaceae	similarity evidence	99% similarity to Acaulium sp. based on E
□	db_xref	taxon:5593	similarity evidence	
□	type_material		similarity evidence	
□	identified_by	Kessy Abarenkov	similarity evidence	

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

Supplementary table 1

Table 1. List of mapped fields available for annotating (ENA feature, ENA qualifier, PlutoF module, PlutoF field, Example)

ENA feature	ENA qualifier	PlutoF module	PlutoF field	Example
source	db_xref (<database:identifier>)	Sequence	Sequence ID	/db_xref=" UNITE:UDB000157"
source	isolation_source (text)	Sequence	Isolation source	/isolation_source="plant leaf"

source	PCR_primers ([fwd_name: XXX,])	Sequence	Forward primer name	/PCR_primers="fwd_name: ITS1F, fwd_seq: CTTGGTCATTTAGAGGAAGTAA, rev_name: ITS4B, rev_seq: CAGGAGACTTGTACACGGTCCAG"
source	PCR_primers (fwd_seq: xxxxx,)	Sequence	Forward primer sequence	/PCR_primers="fwd_seq: CTTGGTCATTTAGAGGAAGTAA"
source	PCR_primers ([rev_name: YYY,])	Sequence	Reverse primer name	/PCR_primers="rev_name: ITS4B, rev_seq: CAGGAGACTTGTACACGGTCCAG"
source	PCR_primers (rev_seq: yyyyy)	Sequence	Reverse primer sequence	/PCR_primers="rev_seq: CAGGAGACTTGTACACGGTCCAG"
source	note (text)	Sequence	Chimeric	/note="This sequence is chimeric"
source	note (text)	Sequence	Low quality	/note="This sequence is of low quality"
source	collection_date (text)	Sequence	Sampling event.Timespan begin	/collection_date="2021-09-28"
source	collection_date (text)	Sequence	Sampling event.Timespan end	/collection_date="2021-09-28/ 2021-09-29"
source	collected_by (text)	Sequence	Sampling event.Collected by	/collected_by="Leho Tedersoo"
source	lat_lon (text)	Sequence	Sampling event.Sampling area.Latitude	/lat_lon="47.94 N 28.12 W"
source	lat_lon (text)	Sequence	Sampling event.Sampling area.Longitude	/lat_lon="47.94 N 28.12 W"
source	country (<country_value>[:<region>][, <locality>])	Sequence	Sampling event.Sampling area.Country	/country="Canada"
source	country (<country_value>[:<region>][, <locality>])	Sequence	Sampling event.Sampling area.State	/country="Canada:Vancouver"

source	country (<country_value>[:<region>][, <locality>])	Sequence	Sampling event.Sampling area.District	/country="Estonia:Harju district"
source	country (<country_value>[:<region>][, <locality>])	Sequence	Sampling event.Sampling area.Commune or City	/country="Estonia:Harju district, Tallinn"
source	country (<country_value>[:<region>][, <locality>])	Sequence	Sampling event.Sampling area.Locality text	/country="Estonia:Harju district, Tallinn, near the harbour"
source	altitude (text)	Sequence	Sampling event.Sampling area.Elevation min.Value	/altitude="320.14 m"
source	altitude (text)	Sequence	Sampling event.Sampling area.Elevation max.Value	
source	altitude (text)	Sequence	Sampling event.Sampling area.Depth min.Value	/altitude="-100 m"
source	altitude (text)	Sequence	Sampling event.Sampling area.Depth max.Value	
source	organism (text); db_xref (<database>:<identifier>)	Sequence	Determination.Taxon name	/organism="Boletus edulis"; /db_xref="taxon:36056"
source	type_material (<type-of-type> of <organism name>)	Sequence	Determination.Typification	/type_material="holotype of Boletus edulis"
source	identified_by (text)	Sequence	Determination.Identified by	/identified_by="Urmas Kõljalg"
source	bio_material ([<institution-code>:[<collection-code>:]]<material_id>)	MaterialSample	Material sample ID	/bio_material=TUE001234
source	host (text)	MaterialSample	Interaction.Taxon	/host="Alnus sp"
source	host (text)	MaterialSample	Interaction.Interacting taxon type	/host="Alnus sp"
source	specimen_voucher ([<institution-code>:[<collection-code>:]]<specimen_id>)	Specimen	Specimen ID	/specimen_voucher=TU<EST>:TUF001234

source	specimen_voucher ([<institution-code>:[<collection-code>:]]<specimen_id>)	Specimen	Subcode	/specimen_voucher=TU<EST>:TUF001234.1
source	culture_collection (<institution-code>:[<collection-code>:]<culture_id>)	LivingSpecimen	Code	/culture_collection=TFC001234
source	culture_collection (<institution-code>:[<collection-code>:]<culture_id>)	LivingSpecimen	Subcode	/culture_collection=TFC001234.1