Enabling Community Curation of Biological Source Annotations of Molecular Data Through PlutoF and the ELIXIR Contextual Data Clearinghouse

Vishnukumar Balavenkataraman Kadhirvelu[‡], Kessy Abarenkov[§], Allan Zirk[§], Joana Paupério[‡], Guy Cochrane[‡], Suran Jayathilaka[‡], Olaf Bánki[|,¶], Jerry Lanfear[#], Filipp Ivanov[§], Timo Piirmann[§], Raivo Pöhönen[§], Urmas Kõljalg[§]

‡ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridge, United Kingdom
§ University of Tartu Natural History Museum, Tartu, Estonia
| Species 2000 / Catalogue of Life, Amsterdam, Netherlands
¶ Naturalis Biodiversity Center, Leiden, Netherlands
# ELIXIR, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridgeshire, United Kingdom

Corresponding author: Vishnukumar Balavenkataraman Kadhirvelu (kadhirvelu@ebi.ac.uk)

Abstract

The advancements in sequencing technologies have greatly contributed to the documentation of Earth's biodiversity. However, for exploring the full potential of molecular resources for biodiversity, there needs to be a good linkage between sequence data and its biological source, contributing to a network of connected data in the biodiversity research cycle. This requires a foundation of well-structured and accessible annotations in the molecular sequence repositories.

The International Nucleotide Sequence Database Collaboration (INSDC), of which the European Nucleotide Archive (ENA) is its European node, holds a large amount of annotations associated with sequence data, relating to its biological source (e.g., specimens in natural history collections). However, for a number of records, these annotations may be incomplete (e.g., missing voucher information), ambiguous or even inaccurate.

Therefore, we have implemented a workflow that allows third-party annotations to be attached to sequence and sample records using two existing services, the PlutoF platform and the ELIXIR Contextual Data ClearingHouse. This work was developed within the scope of the BiCIKL (Biodiversity Community Integrated Knowledge Library) project, which aims to establish open science practices in the biodiversity domain.

PlutoF is an online data management platform that also provides computing services for biology-related research. PlutoF features allow registered users to enter their own data and access public data at INSDC. Users can enter and manage a range of data, as taxonomic

classifications, occurrences, etc. This platform also includes a module that allows the addition of third-party annotations (on material source, taxonomic identification, etc.) linked to specimens or sequence records. This module was already in use by the UNITE community for annotation of INSDC rDNA Internal Transcribed Spacer sequence datasets (Abarenkov et al. 2021). These UNITE annotations are displayed in the National Centre for Biotechnology Information (NCBI) records through links to the PlutoF platform. However, there was the need for an automated solution that allowed third-party annotations to any sequence or sample record at INSDC. This was implemented through the operation of the ELIXIR Contextual Data ClearingHouse (hereafter as Clearinghouse). The Clearinghouse holds a simple RESTful Application Programming Interface (API) to support the submission of additions and improvements to current metadata attributes, such as information on material sources, on records publicly available in the ELIXIR data resources. The Clearinghouse enables the submission of these corrected metadata from databases (such as the PlutoF platform) to the primary data repositories.

The workflow developed is shown in Fig. 1 and consists of the following steps: i) users annotate sequence metadata that is regularly downloaded from INSDC using NCBI's E-utilities; ii) an annotation proposal is created and a verification notification is sent to an assigned reviewer; iii) the reviewer evaluates the annotation proposal and accepts it or rejects it with comments; iv) if the annotation proposal is accepted, the annotated fields that may be mapped to ENA fields are then pushed to the Clearinghouse using their RESTful API. The annotations when received at ENA are then reviewed before being displayed. This workflow is implemented through a web interface in PlutoF, which allows user-friendly and effortless reporting of corrections or additions to biological source metadata in sequence records.

Overall, we expect this tool to contribute to the enrichment of metadata associated with sequence records, and therefore increase the links between the molecular and biodiversity resources, and enable sequencing data to deliver their full potential for biodiversity conservation.

Keywords

third-party annotations, data management, linking data, BiCIKL

Presenting author

Vishnukumar Balavenkataraman Kadhirvelu

Presented at

TDWG 2022

Funding program

BiCIKL project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492. This work was also funded by ELIXIR, the research infrastructure for life-science data.

Grant title

BiCIKL - Biodiversity Community Integrated Knowledge Library

Conflicts of interest

References

- Abarenkov K, Zirk A, Põldmaa K, Piirmann T, Põhönen 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

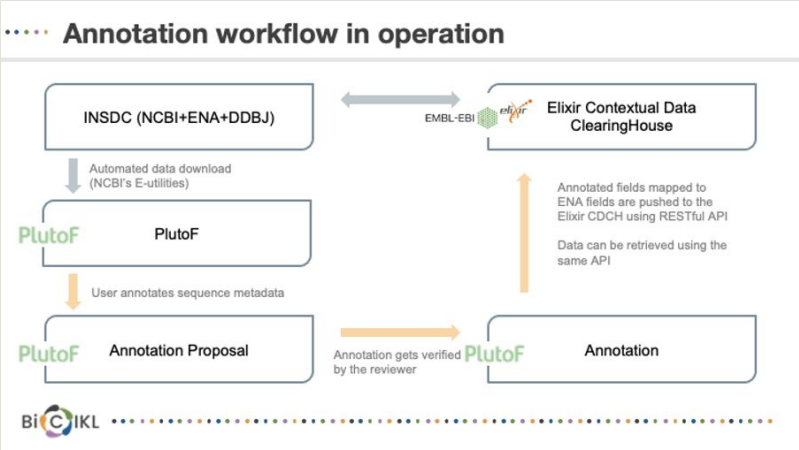

Figure 1.

Workflow for third-party annotations added and verified in PlutoF and submitted to the ELIXIR Clearinghouse.