The ENA Source Attribute Helper: An API for improved biological source data

Vikas Gupta[‡], Joana Paupério[‡], Josephine Burgin[‡], Suran Jayathilaka[‡], Guy Cochrane[‡]

‡ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom

Corresponding author: Vikas Gupta (vikasg@ebi.ac.uk)

Abstract

Metadata management for sequence data is essential for the accurate description of Earth's biodiversity. Within metadata attributes, those that reference the biological sources of sequences and samples and allow linking to the specimen or sample of origin are fundamental for facilitating connections between molecular biology, taxonomy, systematic biology and biodiversity research, increasing the discoverability and usability of data by researchers worldwide.

Sequence data is publicly archived at the International Nucleotide Sequence Database Collaboration (INSDC) that includes the National Centre for Biotechnology Information (NCBI), the DNA Data Bank of Japan (DDBJ) and the European Nucleotide Archive (ENA). Sequences stored at INSDC have associated a considerable range of metadata, including attributes related to its biological source, such as references to natural history collections or culture collections. But, these source attributes are not always submitted or may be incomplete, limiting the association of the sequence records to the original source material, hampering further data connections (e.g., biological data associated with the voucher or species distribution data). Therefore, we have developed the ENA Source Attribute Helper API, a tool that aims to assist users on the submission of accurate attributes referring to the biological source of samples and sequence data. This tool was developed within the scope of BiCIKL (Biodiversity Community Integrated Knowledge Library) (Penev et al. 2022), a Horizon 2020 project which targets building a wide, biodiversity related community for connecting data along the different axes of biodiversity research.

The first version of the tool was designed to support correct annotation of the attributes that identify the source material from which the sample or sequence were obtained, namely /specimen_voucher, /culture_collection, and /biomaterial (INSDC 2021). These attributes follow a Darwin Core Triplet format (Wieczorek et al. 2012), composed of institution code, collection code and the specimen, culture, or material identifier, accordingly.

Since the submission of the biological source attributes to the INSDC may be performed both when data is initially uploaded or on following updates using a variety of tools, we

developed the API as an open source tool that is publicly accessible and may be used as a free-standing service. The API is built using Representational State Transfer (REST) API Architecture and it is designed to use the data available in the NCBI BioCollections (Sharma et al. 2018). NCBI Biocollections is a curated database of metadata for natural history collections, associated with records in INSDC, that includes the institution and collection codes. The API main functions include the querying of the metadata (the API presents both exact matches and similar matches) for the institutions and collections based on the user input, validation of institution and collection codes in the attribute strings provided by the user, and the construction of the attribute string based on the user-provided information. The API does not include the search or validation of the voucher specimen codes.

The API is designed in a way that it can be extended easily for any future enhancements and initially expected to promote and support the submission and any subsequent curation of better structured and more richly described source data. We expect this tool to contribute to better connected biodiversity data and hence provide a stronger foundation to strengthen the value of natural history collections, taxonomic expertise, and biodiversity knowledge.

Keywords

ENA submission tools, validation functions, specimen voucher, culture collection, bio material, NCBI biocollections

Presenting author

Vikas Gupta

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.

Grant title

BiCIKL - Biodiversity Community Integrated Knowledge Library

Conflicts of interest

References

- INSDC (2021) The DDBJ/ENA/GenBank Feature Table Definition. Version 11.1. https://www.insdc.org/feature_table.html#3.2. Accessed on: 2022-6-28.
- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin C, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8: e811360. https://doi.org/10.3897/rio.8.e81136
- Sharma S, Ciufo S, Starchenko E, Darji D, Chlumsky L, Karsch-Mizrachi I, Schoch CL (2018) The NCBI BioCollections Database. Database 2018: article ID bay006. https://doi.org/10.1093/database/bay006
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS One 7 (1): e2971. https://doi.org/10.1371/journal.pone.0029715