



Validation tool open-source software release

Deliverable D8.2

31 October 2022

Authors

Vikas Gupta, Joana Paupério, Josephine Burgin, Suran Jayathilaka, Guy
Cochrane

*European Molecular Biology Laboratory, European Bioinformatics Institute,
Cambridge, United Kingdom*

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: | Validation tool open-source software release |
| Deliverable n°: | D8.2 |
| Nature of the deliverable: | Other |
| Dissemination level: | Public |
| WP responsible: | WP8 |
| Lead beneficiary: | European Molecular Biology Laboratory - European Bioinformatics Institute. |
| Citation: | Gupta, V., Paupério, J., Burgin, J., Jayathilaka, S., Cochrane, G. (2022). <i>Validation tool open-source software release</i> . Deliverable D8.2 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 | 10 October 2022 | European Molecular Biology Laboratory - European Bioinformatics Institute |
| 2.0 | Review | 12 October 2022 | Anton Güntsch Donat Agosti |
| 3.0 | Submission | 25 October 2022 | European Molecular Biology Laboratory - European Bioinformatics Institute |

The content of this deliverable does not necessarily reflect the official opinions of the European Commission or other institutions of the European Union.

Table of contents

| | |
|------------------------------------|----|
| Preface | 4 |
| Summary | 4 |
| List of abbreviations | 4 |
| 1 Background and objectives | 5 |
| 2 ENA Source Attribute Helper tool | 5 |
| 2.1 New version release | 6 |
| 3 API features and implementation | 7 |
| 3.1 Construct and validation flows | 7 |
| 3.2 API endpoints | 7 |
| 3.3 Tools for API access | 8 |
| 4 Graphical user interface | 8 |
| 4.1 Construct tool | 9 |
| 4.1.1 Pre-construct | 10 |
| 4.1.2 Post-construct | 11 |
| 4.2 Validation tool | 12 |
| 4.2.1 Pre-validation | 12 |
| 4.2.2 Post-validation | 12 |
| 5 Future steps | 13 |
| 6 Acknowledgements | 14 |
| 7 References | 14 |

Preface

BiCIKL aims at establishing networks of linked data from the resources of molecular biology, natural history collections, taxonomy and literature. To fully achieve this goal, molecular biology databases need to be populated by well structured biological source metadata that correctly point to the specimens or other material of origin of a sequence or sample. However, users do not always provide comprehensive and well structured source metadata, leading to incomplete or incorrect information on the molecular data repositories. To overcome this issue and help users to provide accurate biological source information, we developed an open access tool, the ENA Source Attribute Helper. This document provides a brief description of the current features of the tool and planned future developments.

Summary

Sample and sequence metadata, that reference correctly the specimens or other material of origin of a sequenced sample and provide links with natural history collections are fundamental for improving the connections between the different areas of knowledge in the life science domain and promoting reusability of data. However, the molecular data repositories don't always hold well structured and comprehensive biological source metadata. Within the scope of project BiCIKL, we developed a tool, the ENA Source Attribute Helper, for aiding users to accurately report biological source attributes of samples and sequences. This version of the tool focuses on the attributes in which specimens, cultures or other materials from which the sequence data was derived, are referred to and uses NCBI Biocollections, a curated database, to obtain the information on the institutions and collections. The tool's main functions comprise the construction of the attribute string based on the user-entered data and the validation of attribute strings provided by the user, while providing additional information about institutions and collections holding the biological material. In this deliverable report we describe the tool's features (in the current release) and expected future developments.

List of abbreviations

| | |
|------|---|
| API | Application programming interface |
| EU | European Union |
| ENA | European Nucleotide Archive |
| GUI | Graphical User Interface |
| NCBI | National Centre for Biotechnology Information |
| SAH | Source Annotation Helper |

1. Background and objectives

Connecting biodiversity data available on the different data resources, such as molecular data, natural history collections, taxonomy and literature, one of the main goals of BiCIKL, is only possible if the available information is accurately provided and structured.

Sequence data and its associated metadata is made available through the International Nucleotide Sequence Database Collaboration (INSDC, Arita et al. 2022), of which the European Nucleotide Archive (ENA, Cummins et al. 2022) is its European node. This infrastructure allows the deposition of enriched metadata associated to sequences, which include biological source attributes, that describe the provenance of the sequence and allow linking to the specimen or material of origin. However, this biological source metadata is user provided, and it is not always complete or unambiguous, preventing often the correct linking of the sequence data to their material of origin.

To overcome this issue, and help the users to provide accurate and linkable biological source attributes of samples and sequence data, we have developed an open source tool, the Source Attribute Helper. This tool will facilitate the submission of well structured biological source information by the users, therefore contributing to increase the discoverability and usability of data by researchers.

2. ENA Source Attribute Helper tool

The ENA Source Attribute Helper is a tool developed to help users provide accurate and complete biological source related metadata. The metadata that refers to the biological source of sequence data is included in samples attributes and, for sequences, is reported in the source feature qualifiers (included in the sequence flat files, INSDC 2021, <https://www.ebi.ac.uk/ena/WebFeat/>). These are referred to hereafter as 'attributes'.

The current version of the tool supports correct annotation of the sequence and sample attributes that identify the specimen, culture, or material from which the sequenced samples are derived, namely /specimen_voucher, /culture_collection, and /biomaterial qualifiers. These attributes follow the Darwin Core Standards (Wieczorek et al. 2012), being the values formatted as a Darwin Core Triplet including the Institution code, collection code and the specimen, culture, or material id accordingly (Table 1).

The tool's main functions are to help users construct and validate the attributes to be submitted as metadata of the sequence or sample, while also providing additional information on the institutions and collections. The tool does not, however, support the search or validation of the voucher identifiers (specimen_id, culture_id or material_id) that need to be obtained directly from the voucher institutions. The search of voucher identifiers will be addressed under the scope of WP7, complementing the current developments. The workflows developed may potentially be integrated in the current system at a later stage.

Table 1: Attributes for the source material supported by the ENA Source Attribute Helper tool. The value format follows the Darwin Core Standards (Wieczorek et al. 2012). Institution code is optional for specimen voucher and bio material, but mandatory for culture collection. Collection code is always optional. When collection code is provided, institution code is mandatory (INSDC 2021). Specimen ID, culture ID, and material ID are mandatory values (from Gupta et al. 2022).

| Qualifier | Definition | Value format |
|--------------------|---|--|
| Specimen voucher | Identifier for the specimen from which the data was obtained | [<institution-code>:<collection-code>:]<specimen_id> |
| Culture collection | Identifier for the culture from which the data was obtained | <institution-code>:<collection-code>:<culture_id> |
| Bio material | Identifier for the biological material from which the data was obtained | [<institution-code>:<collection-code>:]<material_id> |

For obtaining the necessary information on the institutions and collections and for performing string validation, the tool uses the data available in the National Centre for Biotechnology Information (NCBI) Biocollections (<https://www.ncbi.nlm.nih.gov/biocollections>, Sharma et al. 2018). NCBI Biocollections is a curated database of metadata for natural history collections, that are linked to records in INSDC. The database comprises institution and collection codes and the associated URLs where available, among other relevant information. The updated NCBI Biocollections files are fetched manually from the database ftp server and imported into the ENA ElasticSearch datastore.

Biological source attributes are submitted to the ENA during data deposition or in later updates. But there is no single point of entry for metadata submission, as ENA holds several routes for submission (such as RESTful APIs, web interfaces and locally installed command-line tools). Therefore, the ENA Source Attribute Helper was developed as an open source tool that may be used as a free-standing service independently across platforms. The tool includes an Application Programming Interface (API) with several endpoints and a Graphical User Interface (GUI) for a more accessible usage of the tool.

The API main features and implementation are described in section 3 and a detailed description of its design and implementation can be found in Gupta et al (2022). The API is available from a test url (<https://wwwdev.ebi.ac.uk/ena/sah/api/>) and a production url (<https://www.ebi.ac.uk/ena/sah/api/>).

The GUI main features are described in section 4 and the GUI can be accessed through a test url (<https://www.ebi.ac.uk/ena/sah/>) and a production url (<https://www.ebi.ac.uk/ena/sah/>).

The code is available from [GitHub](#) and archived in Zenodo (Jayathilaka & Gupta 2022).

2.1. New version release

Since the initial release of the API (Gupta et al 2022) some updates have been performed including enhancing suggestive search and some minor fixes in the API. Moreover, the GUI was developed. These updates have been released as v.1.1.1, being available at [GitHub](#) and Zenodo (Jayathilaka & Gupta 2022).

3. API features and implementation

3.1. Construct and validation flows

The construct and validation flows of the API are shown in Figure 1 (Gupta et al. 2022). When using the API the user needs to input a code or name (minimum 1 character) for the institution and/or collection and the application will suggest the closest options available, so that the correct option may be selected by the user to proceed with the construction of the qualifier. The API will ensure that the values are aligned with the format definition for each attribute type (/specimen_voucher, /culture_collection, and /biomaterial).

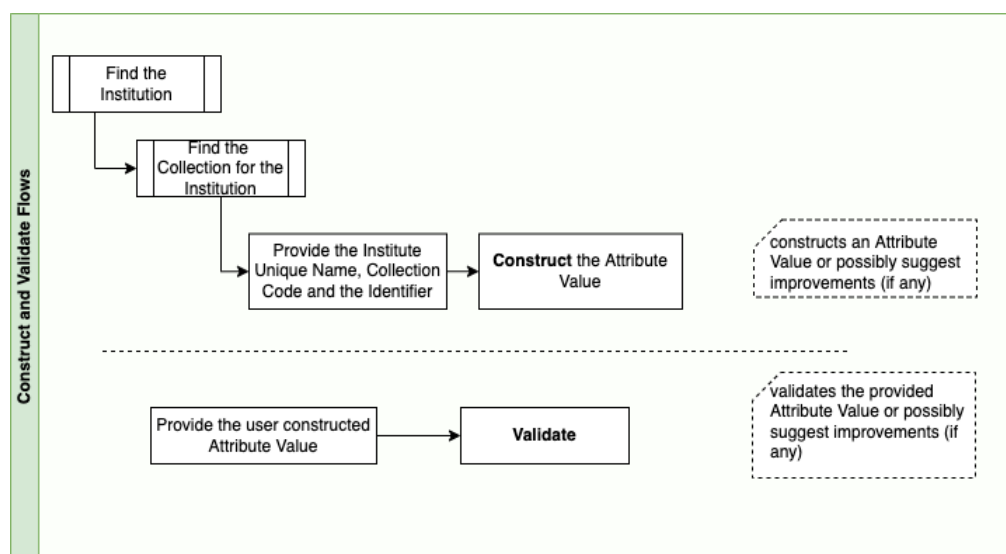


Figure 1: Construct & Validation Flow Diagram for the ENA Source Attribute Helper API (from Gupta et al. 2022).

3.2. API endpoints

The API comprises several endpoints with different functions (Table 2). These include endpoints that allow obtaining the institution code, a specific collection from one institution or the whole range of collections in an institution, as well as endpoints that allow the validation of an attribute string provided by the user and the construction of the attribute string based on values for the different parts (institution, collection and identifier) provided by the user. The endpoint get error-codes is an endpoint that provides users the definitions of the error codes

which may be returned by the system. For more details on the main use cases of the API see Gupta et al (2022).

Table 2: ENA Source Attribute Helper API Endpoints description and success and failure responses (from Gupta et al 2022).

| API Endpoint | Verb | Action | Success | Failure |
|--|------|---|---------|-----------------|
| /institution/{ivalue} | GET | Find an Institution using the institution name or code. If the institution name or code is not fully known, 1 or more characters can be provided. API searches both for exact matches and for partial matches by either institution name or institution code. | 200 OK | 400 Bad Request |
| /institution/{institutionUniqueName}/collection | GET | Gets all collections by the institution's unique name | 200 OK | 400 Bad Request |
| /institution/{institutionUniqueName}/collection/{cvalue} | GET | Gets collection by institution's unique name and collection code. If the collection name or code is not fully known, 1 or more characters can be provided. | 200 OK | 400 Bad Request |
| /validate | GET | Validates the provided attribute string | 200 OK | 400 Bad Request |
| /construct | GET | Constructs the attribute string | 200 OK | 400 Bad Request |
| /error-codes | GET | Gets the error codes definition | 200 OK | 400 Bad Request |

3.3. Tools for API access

There are different modes of access to the API that include:

- any Web Browser
- any scripting/programming language based REST client
- command line tools like cURL and Wget
- testing tools like Swagger UI (user interface) and Postman

4. Graphical user interface

For facilitating the usage of the Source Attribute Helper tool by general users a graphical user interface (GUI) was built. This GUI has two different tools, a **construct tool**, where the user constructs the attribute string for the biological source reference, and a **validation tool**, where the user can validate or enhance a constructed attribute string. The tool is available at <https://www.ebi.ac.uk/ena/sah/>.

4.1. Construct tool

With the Construct tool the user can obtain the correct attribute string to refer to the biological source (voucher or other material) of the sample or sequence data. This tool has a pre-construct display, where the user looks for the institution and collection codes and provides the specimen, culture or material identifier, and a post-construct display, where the constructed value is presented and the user can create a list of constructed attributes. The flow of the construct tool is displayed in Figure 2.

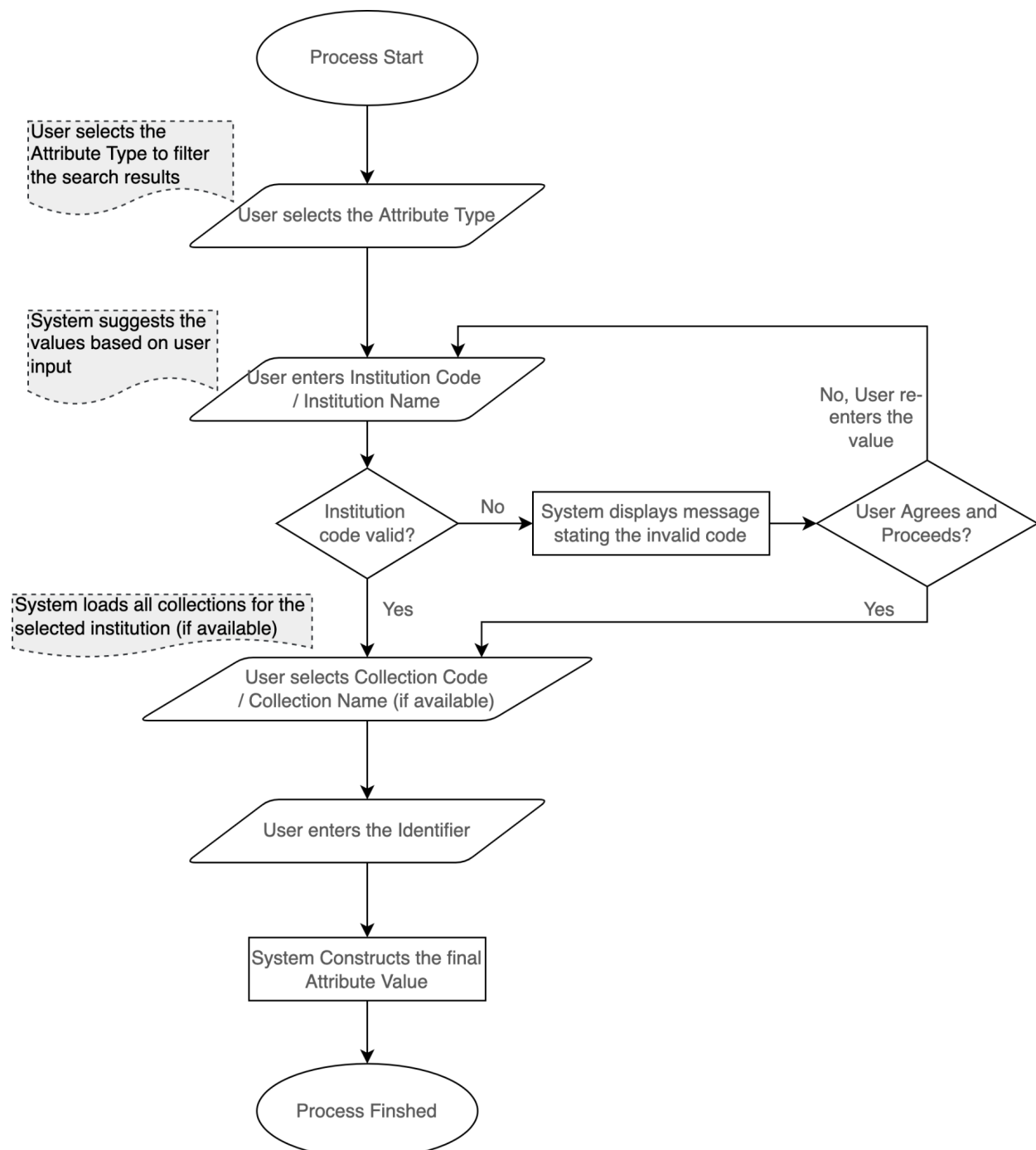


Figure 2: Flow diagram for the frontend construct tool.

4.1.1. Pre-construct

In the pre-construct page the user needs to input some information for the construction of the biological source attribute (Figure 3). The attribute type (specimen voucher, culture collection or biomaterial) needs to be selected so that the tool can fetch only the institutions linked with that attribute type. Then the user may look for the institution code, by starting to type the institution name or code in the ‘Select Institution’ field. When at least one letter is typed the tool matches the search string to both Institution names and codes and suggests in a dropdown possible matches, by displaying the Institutions codes and names. When hovering over the Institutions names, a box with additional information, including the address, is displayed, to facilitate selection of the appropriate code. If the search finds no matches, the message “No Institutions found for the selected criteria” is displayed.

Once the institution is selected the Collection codes are populated in the ‘Select Collection’ dropdown and the user can then select the collection. This field is not mandatory, as not all institutions have registered collections, but it is recommended to be provided if collections are available. Finally the user needs to enter the catalogue number/identifier for the voucher or other material in the ‘Enter Identifier/Code’ field and click on the “Construct Attribute Value” button.

The screenshot shows the ENA Source Attribute Helper (SAH) interface. At the top left is the ENA logo (European Nucleotide Archive). The main heading is "ENA Source Attribute Helper". Below this is a brief description of the tool and its purpose. The interface is divided into two tabs: "Construct" (selected) and "Validate". The "Construct" tab contains instructions and a form with the following fields:

- Attribute Type:** A dropdown menu with "Specimen Voucher (e.g. specimen)" selected. A red asterisk indicates it is mandatory.
- Select Institution:** A text input field with "CMN-Canadian Museum of Nature" entered. A red asterisk indicates it is mandatory.
- Select Collection:** A dropdown menu with "Fish - Fish Collection" selected.
- Enter Identifier/Code:** A text input field with "12345" entered. A red asterisk indicates it is mandatory.

Below the form is a blue button labeled "Construct Attribute Value". A note at the bottom states: "Note: The tool does not include search or validation of the voucher identifier/codes."

Figure 3: Screenshot of the pre-construct page of the ENA Source Attribute Helper GUI.

4.1.2. Post-construct

The attribute value generated is displayed in the ‘Constructed attribute’ section, together with the Attribute type to which it refers to (Figure 4). If links are available at NCBI Biocollections to the institutions and collections pages, these will appear linked in the institute or collection code. There are two additional buttons that allow the user to copy the attribute string (‘Copy’) and to save the value in memory (‘Add to My Constructed Attributes’). When the value is saved a new section at the bottom of the page will be displayed, containing all the saved values by that user (Figure 4). This new section allows also to copy or remove specific values or all the values saved.

ENA
European Nucleotide Archive

ENA Source Attribute Helper

The ENA Source Attribute Helper (SAH) is designed to help users accurately report sequence and sample attributes related to biological sources.

This tool currently focuses on the attributes in which specimens, cultures or other materials are identified, from which the sequence data were derived. It uses curated data from [NCBI BioCollections](#) to obtain unique codes for the institutions and collections holding the vouchers. The accepted formats for these qualifiers are detailed in the [INSDC Feature Table](#).

SAH can be accessed directly/programmatically using the [RESTful API](#)

Construct Validate

You may use this tool to construct in the correct format the attribute string for referring to the biological source of the sequence data.

The Attribute Type (qualifier type) selector constrains the search to institutions that hold the specific type of collection.

Fields marked with asterisk * are mandatory

Attribute Type *
Specimen Voucher (e.g. specimen) *

Select Institution: * ?
CMN-Canadian Museum of Nature

Select Collection: ?
Fish - Fish Collection *

Identifier / Code *
12345

Enter Identifier/Code: * ?

Note: The tool does not include search or validation of the voucher identifier/codes.

Construct Attribute Value

Constructed Attribute:

| Attribute | Attribute Type |
|--------------------------------|------------------|
| CMN:Fish:12345 | Specimen Voucher |

Copy
Add to My Constructed Attributes

My Constructed Attributes: Copy All Remove All

| Attribute | Attribute Type |
|--------------------------------|------------------|
| CMN:Fish:12345 | Specimen Voucher |

Copy Remove

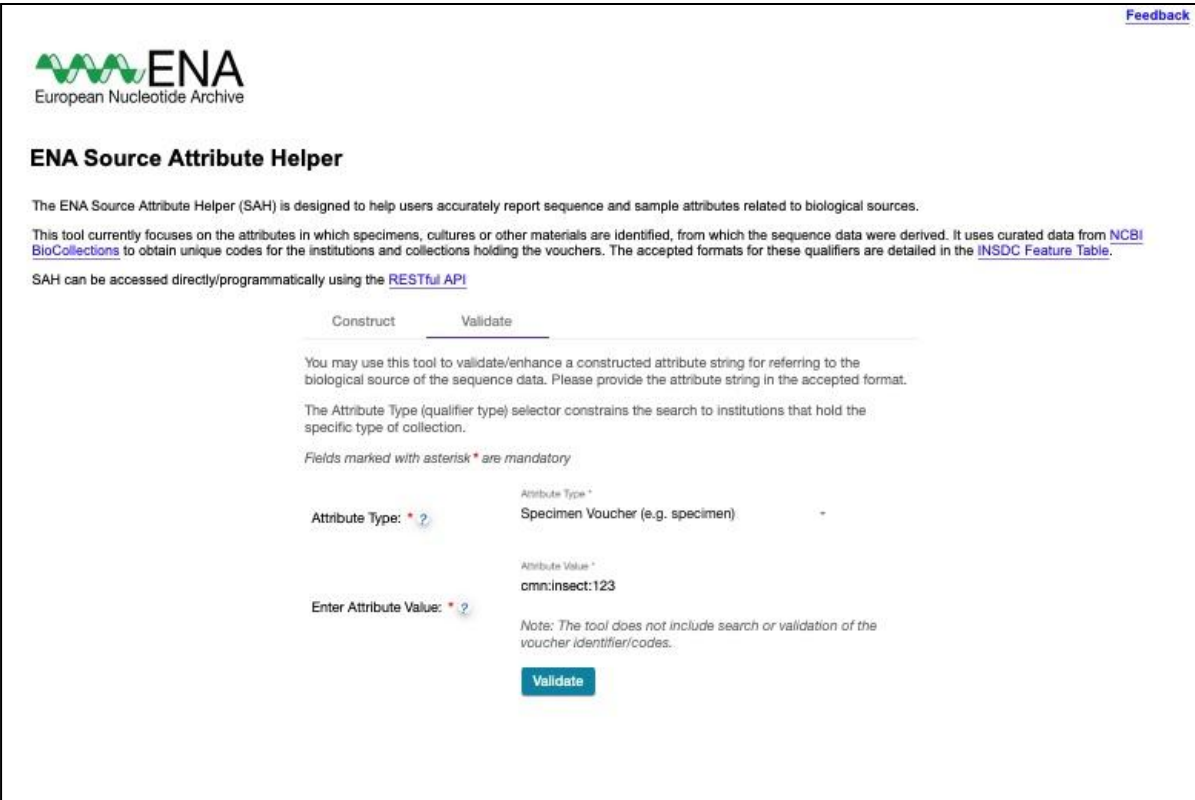
Figure 4: Screenshot of the post-construct page of the ENA Source Attribute Helper GUI.

4.2. Validation tool

The validation tool allows the user to validate and get additional information to a specified attribute string. This tool also has a pre-validation display, where the user inputs the attribute value, and a post-validation display, where the matches found are presented and the user can create a list of validated attribute strings.

4.2.1. Pre-validation

In the pre-validation page the user needs to input information regarding the attribute he wishes to validate (Figure 5). The attribute type (specimen voucher, culture collection or biomaterial) needs to be selected so that the tool can validate the string against the institutions and collections linked with that attribute type. Then, the user needs to enter an attribute string in the 'Enter Attribute Value' field and click on the 'Validate' button.



The screenshot shows the ENA Source Attribute Helper (SAH) interface. At the top left is the ENA logo (European Nucleotide Archive). A 'Feedback' link is in the top right. The main heading is 'ENA Source Attribute Helper'. Below it, a paragraph explains the tool's purpose: 'The ENA Source Attribute Helper (SAH) is designed to help users accurately report sequence and sample attributes related to biological sources. This tool currently focuses on the attributes in which specimens, cultures or other materials are identified, from which the sequence data were derived. It uses curated data from [NCBI BioCollections](#) to obtain unique codes for the institutions and collections holding the vouchers. The accepted formats for these qualifiers are detailed in the [INSOC Feature Table](#). SAH can be accessed directly/programmatically using the [RESTful API](#)'.

The interface has two tabs: 'Construct' and 'Validate', with 'Validate' selected. The 'Validate' section contains the following text: 'You may use this tool to validate/enhance a constructed attribute string for referring to the biological source of the sequence data. Please provide the attribute string in the accepted format. The Attribute Type (qualifier type) selector constrains the search to institutions that hold the specific type of collection. Fields marked with asterisk * are mandatory'.

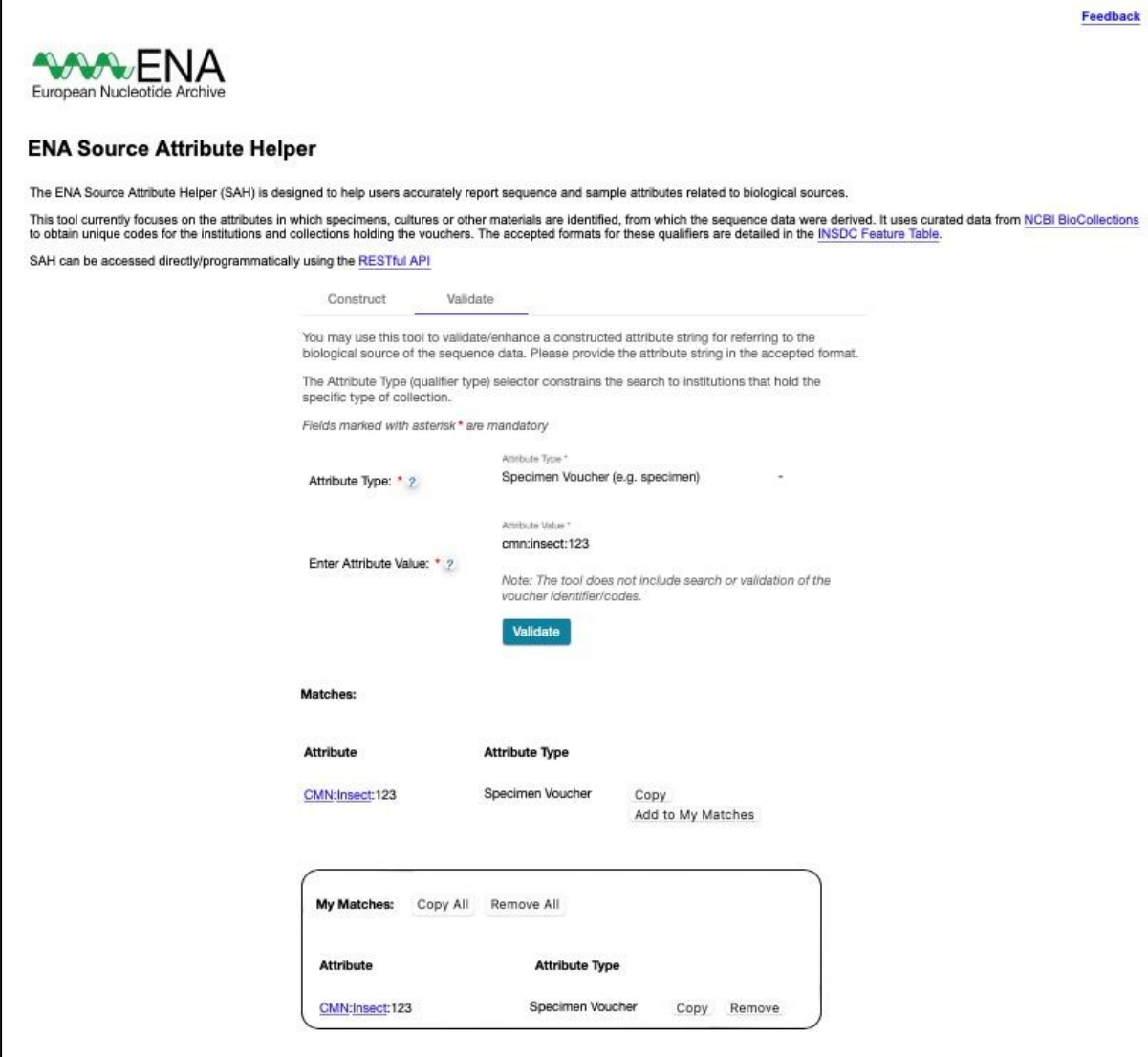
There are two input fields: 'Attribute Type: * ?' with a dropdown menu showing 'Specimen Voucher (e.g. specimen)', and 'Enter Attribute Value: * ?' with a text input field containing 'cmn:insect:123'. A 'Validate' button is located below the input fields. A note at the bottom states: 'Note: The tool does not include search or validation of the voucher identifier/codes.'

Figure 5: Screenshot of the pre-validation page of the ENA Source Attribute Helper GUI.

4.2.2. Post-validation

After the validation of the attribute string a new section appears displaying all possible valid matches according to the input value (Figure 6). These values may contain suggestions of corrections/additions to institutions and collections codes. If links are available at NCBI Biocollections to the institutions and collections pages, these will appear linked in the institution or collection code. As in the post-construct page, two additional buttons that allow

the user to copy the matches ('Copy') and to save the values in memory ('Add to My Matches') will be displayed. If the attribute string input is invalid then an error message will appear. Once the values are saved a new section ('My matches') will be displayed at the bottom of the page, containing all the saved values by that user (Figure 6). This new section allows also to copy or remove specific values or all the values saved.



The screenshot shows the ENA Source Attribute Helper (SAH) interface. At the top left is the ENA logo (European Nucleotide Archive). The page title is "ENA Source Attribute Helper". Below the title, there is a brief description of the tool and its purpose, along with links to "NCBI BioCollections" and "INSDC Feature Table". The main section is titled "Construct" and "Validate". It contains instructions on how to use the tool and a form for entering an attribute string. The form has two main sections: "Attribute Type" and "Enter Attribute Value". The "Attribute Type" section has a dropdown menu with "Specimen Voucher (e.g. specimen)" selected. The "Enter Attribute Value" section has a text input field containing "cmn:insect:123". There is a "Validate" button. Below the form, there is a "Matches:" section with a table showing the entered attribute and its type. The table has two columns: "Attribute" and "Attribute Type". The first row shows "CMN:insect:123" and "Specimen Voucher". There are "Copy" and "Add to My Matches" buttons next to the match. At the bottom, there is a "My Matches:" section with "Copy All" and "Remove All" buttons. Below this, there is a table showing the saved matches. The table has two columns: "Attribute" and "Attribute Type". The first row shows "CMN:insect:123" and "Specimen Voucher". There are "Copy" and "Remove" buttons next to the match.

Figure 6: Screenshot of the pre-construct page of the ENA Source Attribute Helper GUI.

5. Future steps

Future implementations to the Source Attribute Helper tool include:

- implementing an automated flow for retrieving the updated files from the NCBI servers regularly
- Linking the tool to the ENA submission systems
- Addition of Google analytics

6. Acknowledgements

We would like to acknowledge Conrad Schoch, Shobba Sharma and the NCBI taxonomy team for providing information on the NCBI biocollections database and its ftp access. We would also like to thank those who have reviewed and contributed to this document.

7. References

Arita M, Karsch-Mizrachi I, Cochrane G (2021). The international nucleotide sequence database collaboration. *Nucleic Acids Res*, 49: D121–D124. <https://doi.org/10.1093/nar/gkaa967>

Cummins C, Ahamed A, Aslam R, et al (2022) The European Nucleotide Archive in 2021. *Nucleic Acids Research*, 50, D106-D110. <https://doi.org/10.1093/nar/gkab1051>

Gupta V, Paupério J, Burgin J, Jayathilaka S, Cochrane G (2022) ENA Source Attribute Helper: An Application Programming Interface to facilitate accurate reference to biological source data [version 1; peer review: awaiting peer review] (2022). *F1000 Research*, 11: 1042. <https://doi.org/10.12688/f1000research.123934.1>

INSDC (2021) The DDBJ/ENA/GenBank Feature Table Definition. Version 11.1 October 2021. Available at <https://www.insdc.org/submitting-standards/feature-table/>

Jayathilaka S, Gupta V (2022) ENA Source Attribute Helper (v1.1.1). Zenodo. <https://doi.org/10.5281/zenodo.7180841>

Sharma S, Ciufu S, Starchenko E, Darji D, Chlumsky L, Karsch-Mizrachi I, Schoch CL (2018) The NCBI Biocollections Database. *Database* Vol 2018: article ID bay006. <https://doi.org/10.1093/database/bay006>

Wieczorek J, Bloom D, Guralnick R, Blum S, Doring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7(1): e29715. <https://doi.org/10.1371/journal.pone.0029715>