D6.3 Report on the translation module for EU-OS screening centres

Version 1.0 – 2023-11-15

for

H2020-INFRADEV-2018-1

(Development and long-term sustainability of new pan-European research infrastructures)

Research and Innovation Action (RIA)

Action Acronym:

EU-OPENSCREEN-DRIVE

Action Full Title:

“Ensuring long-term sustainability of excellence in chemical biology within Europe and beyond”

Grant Agreement No.: 823893

Dissemination level: public

Document Properties:

Deliverable	D6.3 Report on the translation module for EU-OS screening centres
Responsible partner(s)	EU-OS, IMG
Author(s)	Katja Herzog (EU-OS), Ctibor Skuta (IMG)
Due date	31/08/2023

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

1 Introduction

EU-OPENSCREEN (EU-OS) was established in 2018 to enable and support molecular and cell biologists as they implement their own projects in chemical biology and early drug discovery. As a distributed research infrastructure, EU-OPENSCREEN works with over 30 academic partner institutes across 10 European countries. Twenty-two screening platforms and 11 medicinal chemistry groups offer their expertise and instrumentation for the development of novel chemical probes in collaboration with the broader biology community.

The scientific screening (user) projects conducted at the EU-OS partner sites use the jointly used compound collection of EU-OS, which contains commercial compounds as well as “academic” compounds submitted by chemists from all over Europe. All structural data of the compounds and primary screening data are made available to the scientific community in the open-access EU-OS European Chemical Biology Database (ECBD; https://ecbd.eu/), which has been developed and is hosted by the EU-OS database partner site IMG in Prague. The data are essential for researchers who are interested in understanding the structure–activity relationships of their compounds or repurposing existing drugs for new targets.

The goal of work package 6 is to improve and harmonize the process of data management across the partner site network by providing new tools and processes aligned with the FAIR data guiding principles, which were developed to improve the Findability, Accessibility, Interoperability, and Reuse of scientific data and ensure that all participating screening centers can adopt such standards directly into their operating systems. To do so, the aim was to offer substantial support for the use of standards at every stage of data capture and transfer. This also implied the development of a translation module (D6.3) for EU-OS screening centers, which was meant to manage the bidirectional translation of data formats from the local standards and data structures at EU-OS screening sites into the ECBD standard format.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

2 Report on the translation module

To achieve a high level of reusability and reproducibility of the experimental data uploaded to ECBD, data must be described in great detail using established ontologies/vocabularies, utilizing common IDs and data formats, and complemented with the experimental protocols and other relevant documents/attachments. The scheme of the data upload process is shown in Figure 1.

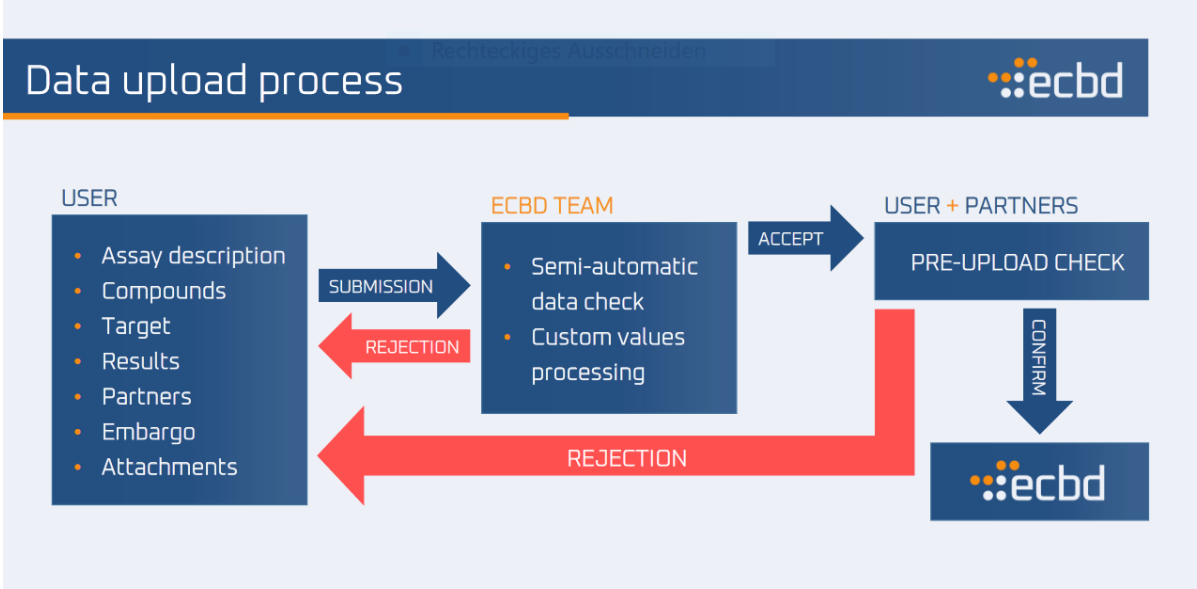

Figure 1 Data upload process into the ECBD.

Since all EU-OS partner sites use the same compound collection (and its subsets), the identification of used samples can be simply resolved through the centrally assigned compound EU-OS IDs (e.g., EOS1); these IDs are also used to link to the compounds (e.g., https://ecbd.eu/compound/EOS1).

The most complex and time-demanding part of the upload is the metadata description utilizing various ontologies (in the case of ECBD: *BioAssay ontology, BRENDA Tissue Ontology, Cellosaurus, NCBI Taxonomy, Reactome Pathway Ontology*). Each of these represents a complex description tool: in some cases, containing hundreds of thousands of terms, and each has a unique website and user interface with a different level of user accessibility/usability. To streamline use and ensure a consistent metadata description process, it is necessary to integrate these ontologies and clarify which of their parts (subsets) should be used within each specified metadata field.

We achieved this aim by creating a user-friendly and intuitive upload interface that incorporates all mentioned metadata sources (ontologies, vocabularies, standard identifiers) and context-dependent metadata components to simplify the data description (Figure 2, Figure 3, Figure 4). This streamlining significantly reduces the complexity of uploaded experimental data, enabling all partner screening centers to export the data in a simple specified format (Figure 5).

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

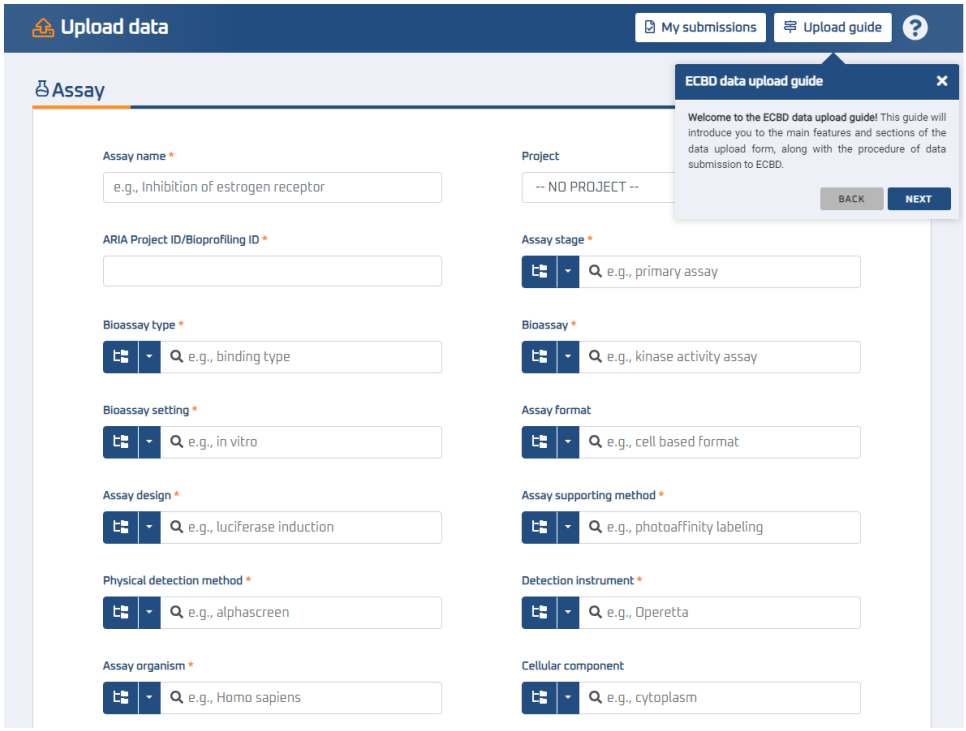

Figure 2 ECBD data upload interface. Most metadata description fields are connected to the relevant part of the selected ontologies.

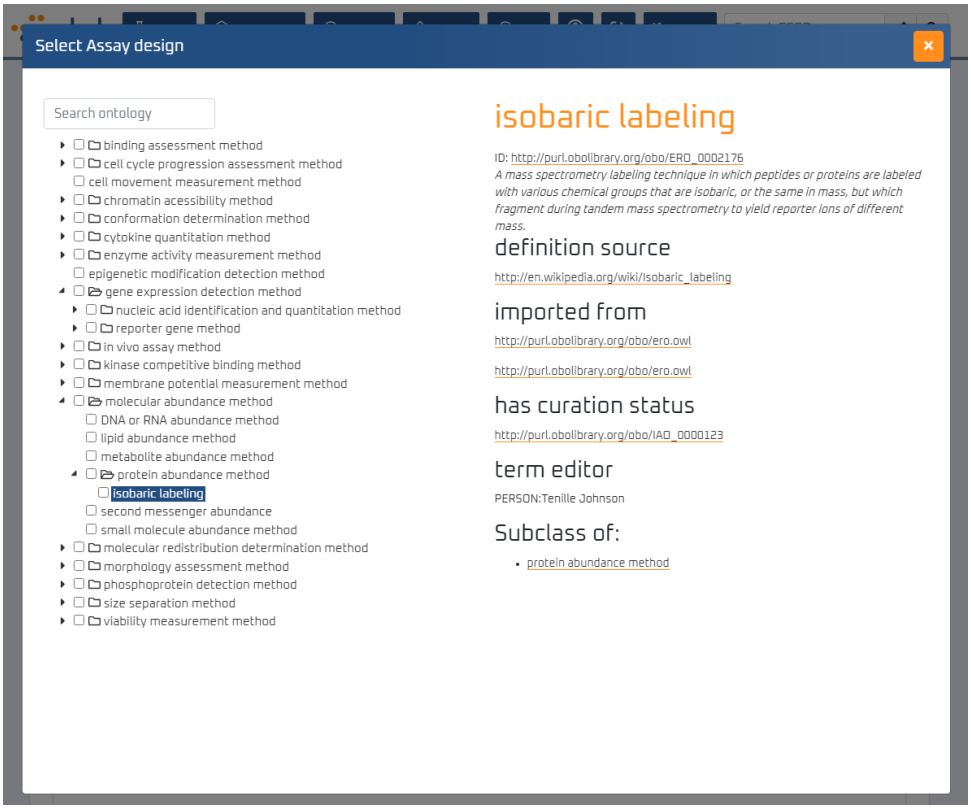

Figure 3 The selection of a term in a specified branch of BioAssay ontology associated with the metadata field for the description of an assay design.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

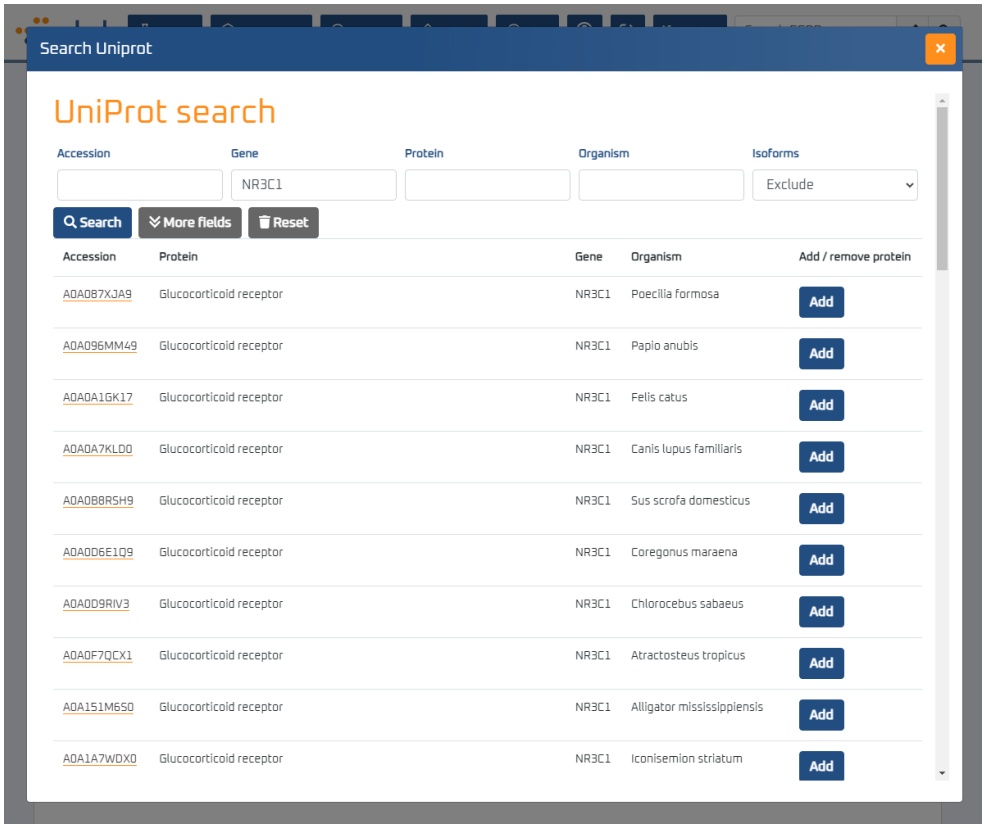

Figure 4 The integrated UniProt search used for the specification of target protein(s).

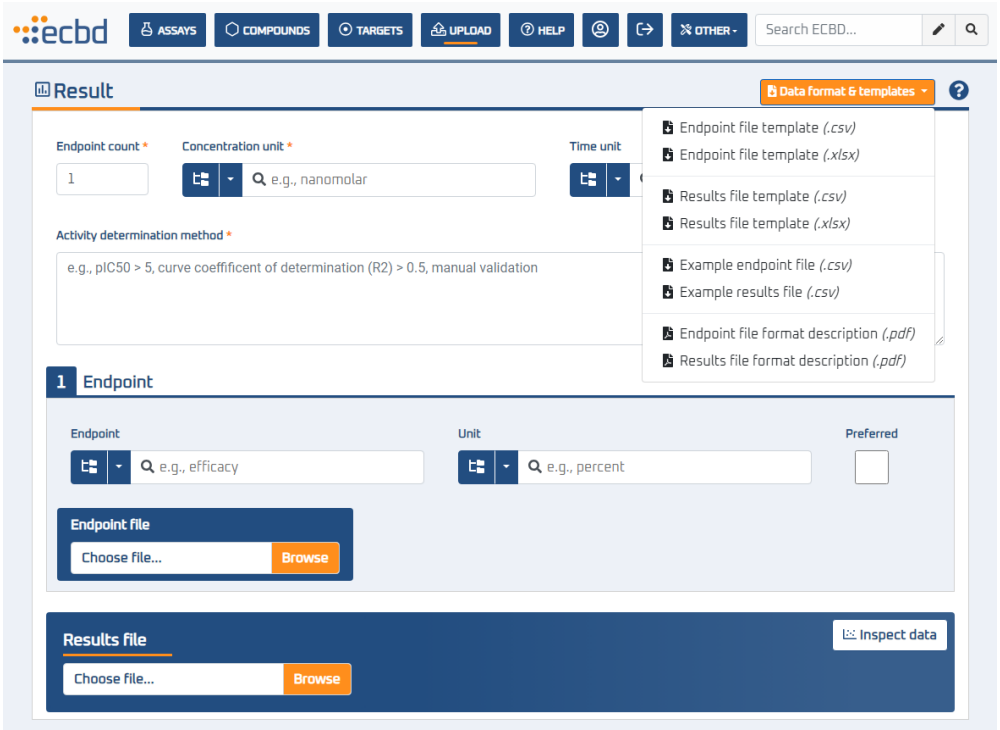

Figure 5 The result (experimental data) section of the ECBD upload form, including data file templates, examples, and format description.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893

Initial plans to implement programmatic translation modules to facilitate data upload from individual screening centers to the ECBD were abandoned during development, as the central upload interface was designed to minimize reliance on additional data translation tools. A detailed description of the data upload process can be found in D6.2: Report on the implementation of enhanced transfer procedures for data from EU-OS partner sites.

The decision to avoid programmatic translation modules aligns with our goal of minimizing system complexity. Reducing the number of system components enhances robustness, ease of maintenance, and overall efficiency.

3 Conclusion

In conclusion, the decision to not implement programmatic translation modules reflects a strategic shift towards simplification and efficiency. The development of a centralized, intuitive upload interface that effectively integrates various description sources has rendered these modules redundant. This approach not only streamlines the data upload process but also ensures consistency and ease of use for all participating screening centers. By focusing on reducing system complexity, the project significantly enhances the robustness, maintainability, and overall efficiency of data management across the partner site network, aligning with the FAIR data guiding principles.

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under the grant agreement No. 823893