An evaluation of EDAM coverage in the Tools Ecosystem and prototype integration of Galaxy and WorkflowHub systems

Lucie Lamothe, Jennifer Rugaard Bregndahl Jensen, Hans Ienasescu, Ove Johan Ragnar Gustafsson, Alban Gaignard, Dmitry Repchevsky, Radka Svobodová, Tomáš Raček, Matej Antol, Magnus Palmblad, et al.

BioHackrXiv

An evaluation of EDAM coverage in the Tools Ecosystem and prototype integration of Galaxy and WorkflowHub systems

Lucie Lamothe[1], **Jennifer Rugaard Bregndahl Jensen**[1], **Hans Ienasescu**[1], **Ove Johan Ragnar Gustafsson**[6], **Alban Gaignard**[11], **Dmitry Repchevsky**[8, 9], **Radka Svobodová**[4, 5], **Tomáš Raček**[4, 5], **Adrián Rošinec**[4, 12], **Matej Antol**[12], **Magnus Palmblad**[3], **Matúš Kalaš**[10], **and Hervé Ménager**[1, 2]

1 Institut Français de Bioinformatique, Evry, F-91000, France **2** Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France **3** Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, The Netherlands **4** CEITEC-Central European Institute of Technology, Masaryk University, Kamenice 5, 602 00, Brno, Czech Republic **5** National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00, Brno, Czech Republic **6** Australian BioCommons, University of Melbourne, Victoria, Australia **7** National Life Science Supercomputing Center, Technical University of Denmark, Denmark **8** Spanish National Bioinformatics Institute (INB-ISCIII), Barcelona, Spain **9** Barcelona Supercomputing Center (BSC), Plaça Eusebi Güell, 1-3 Barcelona, Spain **10** Department of Informatics, University of Bergen, Norway **11** L'Institut du Thorax, University of Nantes/CNRS/INSERM, France **12** Institute of Computer Science, Masaryk University, Šumavská 15, 60200 Brno, Czech Republic

BioHackathon series:
BioHackathon Europe 2022
Paris, France, 2022
Project 25

Submitted: 15 Feb 2023

Published by BioHackrXiv.org

Introduction

The Tools Ecosystem is a centralized repository for the open and transparent exchange of metadata about software tools and services in bioinformatics and life sciences. It serves as a foundation for the sustainability and interoperability of the diverse Tools Platform services: bio.tools (Ison et al., 2019), BioContainers (Veiga Leprevost et al., 2017), OpenEBench (Capella-Gutierrez et al., 2017), Bioconda (Grüning et al., 2018), WorkflowHub (Goble et al., 2021), usegalaxy.eu (Community, 2022a). It also includes a number of related resources outside of the ELIXIR Tools Platform (e.g. Debian Med, biii.eu).

Here we report the results of a project started at the BioHackathon Europe 2022. Its goals were to:

1. Cross-compare and analyze the metadata centralized in the Tools Ecosystem, including coverage of the EDAM ontology (Ison et al., 2013), and
2. Explore methods for connecting tools used in registered Galaxy workflows (i.e. WorkflowHub entries) to the annotations available in bio.tools.

This report is separated into three sections. In the first, we present the results of the two analyses described above. The second section details the methods we used, and finally we discuss potential perspectives for both improved monitoring and curation of the Tools Ecosystem metadata and EDAM, as well as for improving the connectivity and integration between elements of the Ecosystem (i.e. bio.tools, WorkflowHub) and platform services that make use of this Ecosystem (e.g. Galaxy (Sloggett et al., 2013)).

Results and discussion

Semantic annotation of bio.tools entries of EDAM

Here we assess the completeness of the annotation of bio.tools entries with EDAM concepts. The two figures 1a and 1b represent the respective proportions of entries annotated with EDAM topics, operations, data and formats.

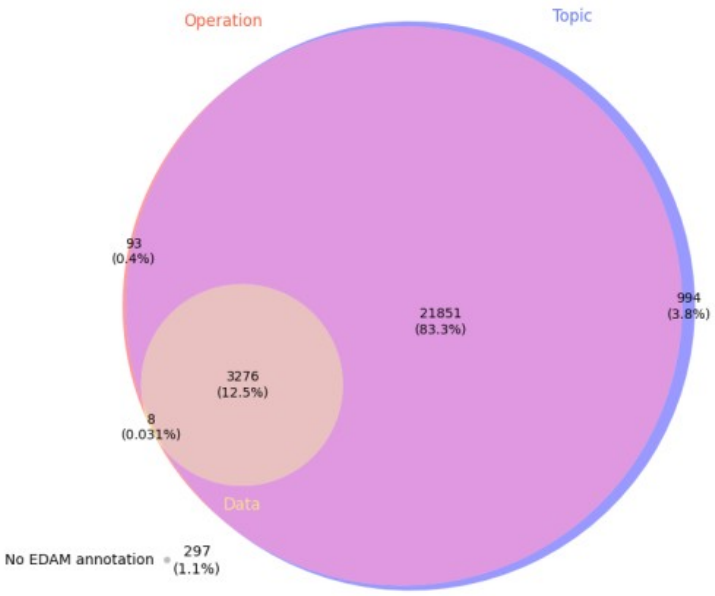

Figure 1a: bio.tools entries annotation represented as a Venn diagram. Each set here represents the proportion of entries annotated with EDAM topics, operations and data. Overlap areas indicate the proportion of bio.tools entries annotated with e.g. both topics and operations, or topics, operations, and data.

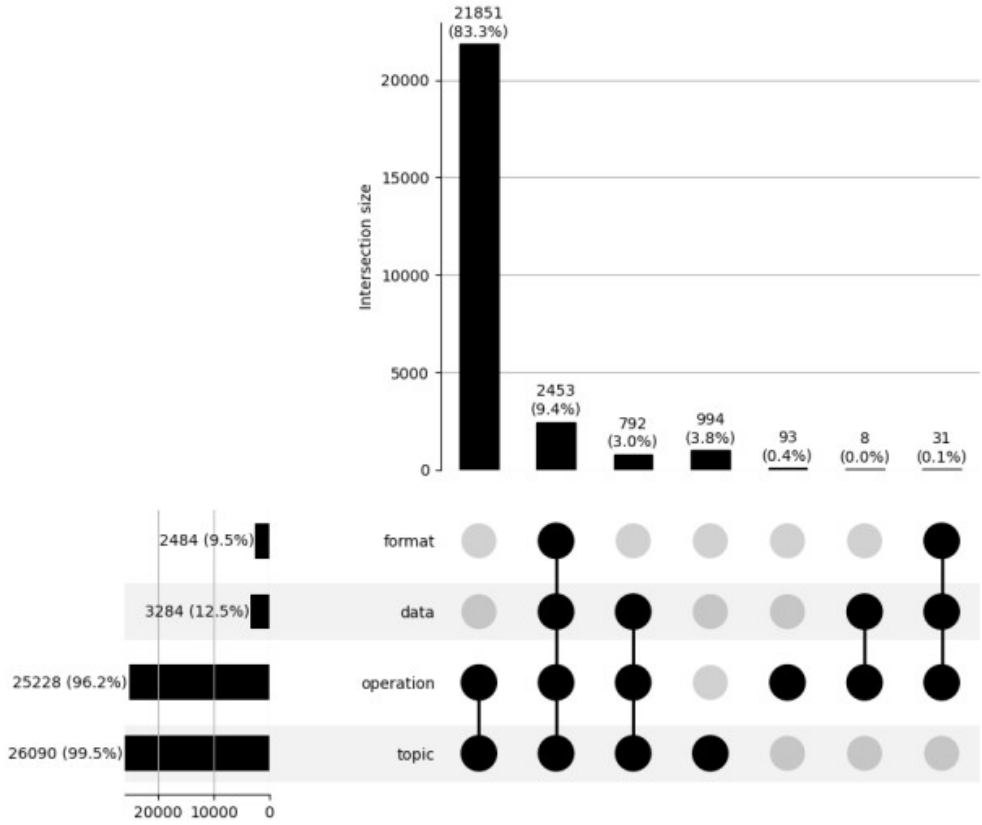

Figure 1b: bio.tools entries annotation represented as an UpSet plot. Each line represents the proportion of entries annotated with EDAM topics, operations, data and formats. Columns indicate the proportion of bio.tools entries for each combination of intersections (e.g. proportion of entries annotated with topics and operations).

These results show that most entries (83.3%) are annotated with topics and operations, and another significant portion (9.4%) also includes data and format concepts. Only a very small proportion of the entries (1.1%) do not include any EDAM annotation. The explanation for the important proportion of annotated entries is probably that EDAM topics and operations can be specified either by human contributors and curators, but also as the result of an automated text-mining process (Jaaniso, 2016). This level of annotation, with 96.2% of the tools including operation annotations, means that for almost all entries the scientific functions of the tools is described.

EDAM usage in bio.tools

The analyses in this section assess the usage of EDAM concepts in bio.tools for each of its four sections. We first evaluate the proportion of the EDAM concepts used to annotate entries, and then relate their usage to their suitability for annotation: some EDAM concepts, obsolete or auxiliary, should not be used for annotations.

This usage analysis has been performed with two versions of EDAM: the last stable release (1.25) currently used in bio.tools, and the current development version (commit hash: 38f21c1edf839efa5d957395f9495562cc7dc1f8). The comparison of these two versions allows to foresee the impact of the future release of EDAM on bio.tools annotations.

The goal of this approach is mainly to provide a first assessment on the fitness EDAM and bio.tools, i.e. evaluate whether the space of scientific concepts available in EDAM is adapted to the semantic description of bio.tools entries, and highlight some curation priorities.

EDAM Concept categories

Each of the different EDAM sections analysed here are separated into the following categories:

- *Obsolete concepts*: concepts that are judged to be obsolete in the current version. They are marked as deprecated but can still be referenced with their URI.
- *Auxiliary concepts*: concepts which are usually placeholders to organize other sub-concepts and should not be used for annotation purposes. Technically speaking, these are identified using the "notRecommendedForAnnotation" property.
- *Valid concepts*: concepts which are valid for annotation (neither deprecated, nor auxiliary).
- *New (valid) concepts*: concepts added to the ontology in the current development version (future 1.26).

Topics:

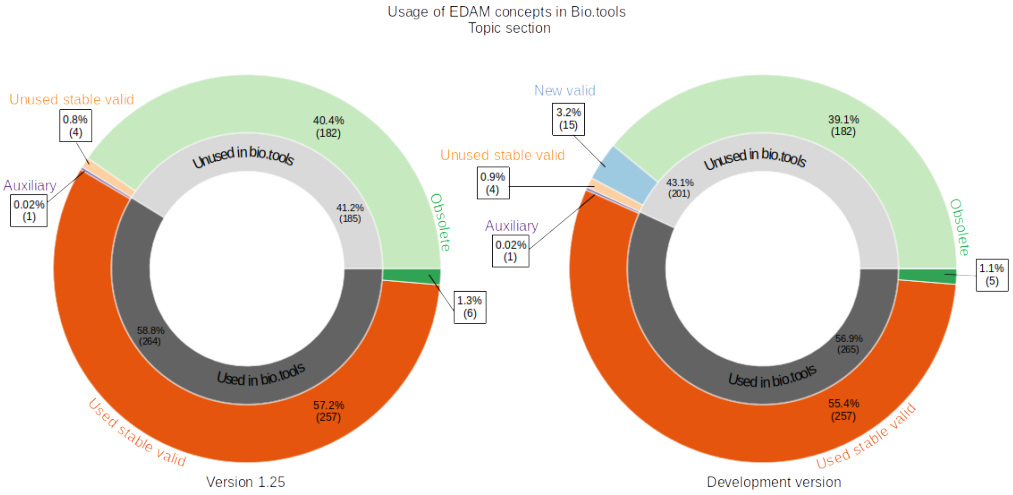

Figure 2: Usage of EDAM Topic concepts in bio.tools

As shown in Figure 2, EDAM topics usage in bio.tools reveals no major anomaly, most of the *valid topics* from version 1.25 being used in EDAM, whereas *obsolete* concepts are only marginally present in annotations. This could optimistically be interpreted as the indication of the fitness for the use of EDAM topics in bio.tools. However, this result doesn't guarantee that the topics section is extensive or precise enough, as some concepts can be used as default for lack of a better one.

Operations:

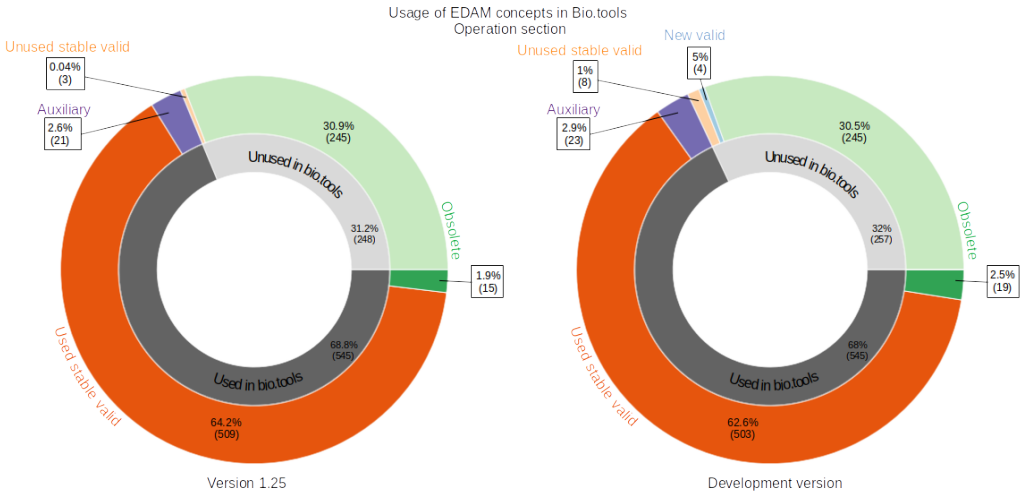

Figure 3: Usage of EDAM Operation concepts in bio.tools

Figure 3 displays the same tendencies for EDAM operations, with an overall wide usage of *valid operations* while *obsolete* ones are mostly unused. However, a frequent issue is the use of *auxiliary* concepts. Although such annotations are probably accurate (i.e. consistent with the scientific function performed by the tool), the usage of such concepts is usually discouraged as they are either too broad or reserved for internal structural purposes. The status of some auxiliary concepts from EDAM might however be revised, if such concepts are the most relevant for annotation.

Data and Formats:

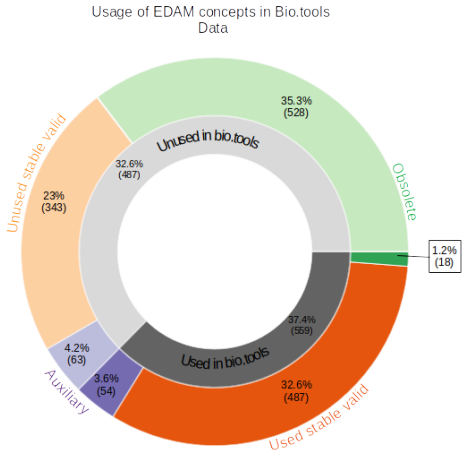

Figure 4: Usage of EDAM data concepts in bio.tools *(only one plot is shown as no new data concepts have been added to the current development version of EDAM)*

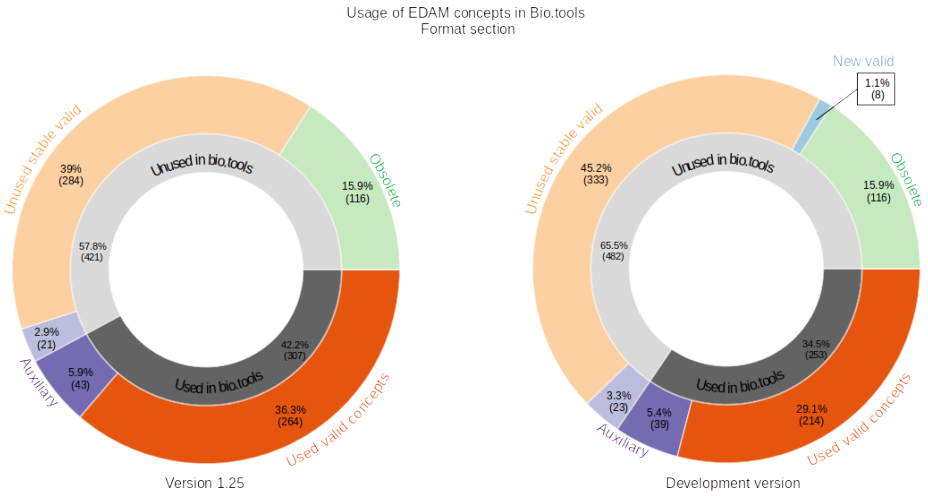

Figure 5: Usage of EDAM format concepts in bio.tools

In contrast with the usage of topics and operations, a large proportion of data (23%) and formats (45.2%) are not used in bio.tools. Further investigation will be needed to determine this low usage is a consequence of annotation issues, or whether it is justified. Potential explanations include:

- a lower level of bio.tools EDAM annotation with data and formats, probably linked to the lack of data/format annotations in text-mining created entries.
- a scope difference between EDAM data and formats and bio.tools, as e.g. a number of EDAM formats might not be relevant in the scientific scope of bio.tools.
- the irrelevance of "legacy" formats related to life sciences technologies and tools which are now deprecated. It is important to note that such legacy file formats cannot be automatically deprecated as they are still used by older tools and can still be found in databases. This may explain the lack of usage of some valid formats in bio.tools.

Mapping between WorkflowHub and bio.tools

Here, we sought to explore whether bio.tools identifiers, and by extension EDAM annotations, could be integrated with WorkflowHub entries for Galaxy workflows. WorkflowHub allows developers to register workflows, each of which are composed of one or more software tools. The connection to bio.tools is clear, and one can imagine a scenario where a workflow registered in WorkflowHub:

1. Has component tools automatically extracted (as is the case currently for Galaxy);
2. Each tool has a bio.tools identifier;
3. This identifier allows WorkflowHub to import and present bio.tools annotations in workflow entries;
4. WorkflowHub can filter workflows based on both EDAM terms and bio.tools identifiers (currently available functionality); and
5. bio.tools can perform the reverse operation and import metadata about workflows that use specific bio.tools entries

To link WorkflowHub and bio.tools entries, all Galaxy workflows from WorkflowHub (https://workflowhub.eu/workflows) were accessed via API and, where metadata was available (82 out of 129 total workflows), a map was created between the individual tool steps in these workflow entries and Galaxy tool identifiers. This ultimately provided a list of workflow steps, mapped to identifiers from WorkflowHub, Galaxy and bio.tools.

Results

Metric	Value
No. of tools WITH a bio.tools ID	513
No. of tools without a bio.tools ID	302
Total no. of tools for all workflows	815
Total no. of workflows used	82

The results of the mapping revealed that for 815 tools used across 82 workflows, 513 tools had a bio.tools identifier (63%). Note that the absence of a mapped bio.tools identifier does not mean that it does not exist. It is also possible that the identifier exists but that it still needs to be added to the Galaxy tool metadata. For example, `hifiasm` is used by the workflow `PacBio HiFi genome assembly using hifiasm` (Price & Farquharson, 2022). This tool has a bio.tools identifier (https://bio.tools/hifiasm) which could be added to the Galaxy tool wrapper.

The table below shows the WorkflowHub identifier (with links) and the **unique** bio.tools identifiers extracted from 10 example workflows.

WorkflowHub ID	bio.tools IDs
138	bedtools, bx-python, bcftools
221	hifiadapterfilt, bandage
395	cutadapt, bowtie2, samtools, bedtools, macs, multiqc
397	cutadapt, bowtie2, samtools, macs, multiqc
398	cutadapt, bowtie2, samtools, macs, multiqc
399	cutadapt, bowtie2, samtools, bedtools, seqcode, macs, multiqc
400	cutadapt, star, multiqc, cufflinks, bedtools
403	quast, busco, merqury
406	nanoplot, minimap2, Racon, unicycler, miniasm, bandage, staramr
407	bbmap, shovill, bwa, pilon, mob-suite, SISTR

Methods

To facilitate the analysis of the data extracted from the Tools Ecosystem and other resources, we integrated them in a local SPARQL endpoint, using the GraphDB commercial software, a solution that enables the querying of RDF resources. The various resources uploaded to a GraphDB-based SPARQL endpoint include:

- the EDAM ontology (Ison et al., 2020), available in its development version on the EDAM GitHub repository.
- the bio.tools contents (Ison et al., 2019), available on the Tools Platform Ecosystem git repository as a Turtle-formatted BioSchemas (Gray et al., 2017) file.
- the workflow data extracted from WorkflowHub and Galaxy, and formatted as well in a BioSchemas format.

The analysis of the data is performed using SPARQL queries, which are performed using a number of Jupyter (Kluyver et al., 2016) notebooks. The various results are visualized using python libraries such as matplotlib (Hunter, 2007).

Mapping between WorkflowHub and bio.tools

The functions for mapping between WorkflowHub and bio.tools:

1. Access the Galaxy Australia and Galaxy Europe APIs to extract both Galaxy specific tool identifiers and bio.tools identifiers, where available (Community, 2022b);

2. Access the entire WorkflowHub registry via API (https://workflowhub.eu/workflows.json), filtering for Galaxy workflows only;
3. Collect all available workflow metadata from the WorkflowHub API;
4. For each workflow, extract all workflow step numbers and Galaxy identifiers (where these are documented, N = 82 of N = 129 workflows total);
5. Map and extract Galaxy tool identifiers for each workflow step, WorkflowHub identifiers (i.e. unique number), bio.tools identifiers for each workflow step, and the workflow step numbers; and
6. Convert mapped identifiers to `*.ttl` format

The functions described are available here. Code is based on previous work for ToolFinder (Gustafsson et al., 2021). See also https://github.com/AustralianBioCommons/australianbiocommons.github.io.

Perspectives

The results presented here represent a first approach to building a knowledge base that integrates data from the various Tools Ecosystem components. Based on these results, we plan to develop further a series of goals, which we describe below:

- *Publication of an open SPARQL endpoint* Improving on the results presented above, we will provide a publicly available SPARQL endpoint using an opensource software. This endpoint will be automatically updated by the Tools Ecosystem CI workflows. It will allow any user to query a dataset that will include not only EDAM, bio.tools, and WorkflowHub, but many other resources, and our teams of maintainers to run queries periodically if needed. It could also be used to optimize the Caseologue EDAM CI tool.
- *EDAM and Bio.tools analysis* With this work we created and tested a valid work environment for analysing the EDAM ontology and the Bio.tools registry. In the same fashion as we plan on lifting up the POC level knowledge base up to an open public SPARQL endpoint, we plan on enriching the analyses to provide an exhaustive set of results that monitor the evolution and consistency of the Tools Ecosystem, for both maintainers and the Tools Ecosystem users. These will be run regularly by the Tools Ecosystem CI.
- Thanks to our first query base we can easily imagine building on them to be able to evaluate and enhance metadata from the whole Ecosystem. Furthermore, we could to do so automatically to track progress using appropriate graphic visualizations, for the benefit of both maintainers and Tools Ecosystem communities and users.

This first round of analysis left us with some identified curation tasks for both EDAM and bio.tools that should be investigated in the future, as discussed in the Results and Discussion. Moreover, some potential enhancement were identified, such as the possibility to track the provenance of bio.tools entry annotations, which would allow to identify more easily e.g. whether they were generated by text-mining tools, or by human contributors.

This work also raises the question of the handling of deprecation in bio.tools. As of 2022, deprecated terms are not removed from the bio.tools annotation. This cannot be properly handled automatically and would be too time consuming manually. However, the total or partial automation of annotation updates could make use of "replacedBy" or "consider" EDAM properties which are used for deprecated concepts.

To resolve the lack of annotation of bio.tools entries using data and format concepts from EDAM, a perpective could be to improve bio.tools interface. All operations in EDAM are linked to a data via a "has_input" and a "has_output" relation (using inferences from parent concepts). For each operation added to the tool's annotation, input and output data would be suggested to the curator/author based on EDAM relation "has_input", "has_output". The same could go for suggestion of format based on its relation with data as 533 formats (over

the 619 valid formats) are related to a data with the "is_format_of" relation. For human user we could also have in the bio.tools interface a "suggested input/output" that would be displayed on the tool page but clearly identifed as an unverified annotation. The enhancement of complete EDAM annotation could lead to automatic workflow generation using the whole Ecosystem metadata.

Mapping between WorkflowHub and bio.tools

In the previous section we detailed new approaches for evaluating and improving EDAM annotations in bio.tools: the other side of this coin is to ensure that other parts of the Ecosystem can share and benefit from these improvements. One of these parts is computational workflows that are registered in WorkflowHub. Each of these workflows represent a significant investment of researcher time and expertise: ideally these workflows would be able to draw directly on the wealth of metadata and EDAM ontology annotations stored by a registry like bio.tools, with minimal additional input of effort from a workflow developer. The prototype map described by this project is incomplete, with 37% of Galaxy workflow tools missing a bio.tools identifier. Likely causes include either missing annotations in the Galaxy tool wrappers or missing bio.tools registry entries. However, the potential value is already clear: a potential next step is for the map functions between WorkflowHub and Galaxy to be productionised by the Tools Ecosystem, such that bio.tools is able to access all Galaxy workflows on WorkflowHub (N = 129 in January 2023) that make use of bio.tools entries, and WorkflowHub is able to access tool components of its workflows as well as bio.tools registry metadata. Synchronisation in this manner will give each platform the opportunity to further improve the experience of users that contribute to and maintain a FAIR software ecosystem.

Many thousands of Galaxy workflows exist globally. With automated integration, users of WorkflowHub will be able to intuitively navigate the growing set of Galaxy workflows based on their tool of choice, topic, or software operation.

Code availability

The code described to run the analyses and obtain the results presented here is freely available on a dedicated GitHub repository. The data collected are also freely available on the Tools Ecosystem main repository and on the EDAM repository.

Acknowledgements

This work was funded/supported by ELIXIR, the research infrastructure for life-science data. This work was supported by the Australian BioCommons which is enabled by NCRIS via Bioplatforms Australia funding. The authors wish to thank the WorkflowHub, Bioschemas, and Galaxy project teams for their technical help and the fruitful discussions that led to these results.

References

Capella-Gutierrez, S., Iglesia, D. de la, Haas, J., Lourenco, A., Fernández, J. M., Repchevsky, D., Dessimoz, C., Schwede, T., Notredame, C., Gelpi, J. L., & Valencia, A. (2017). Lessons learned: Recommendations for establishing critical periodic scientific benchmarking. *bioRxiv*. https://doi.org/10.1101/181677 **[cito:usesDataFrom]**

Community, T. G. (2022a). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, *50*(W1), W345–W351. https://doi.org/10.1093/nar/gkac247 **[cito:usesDataFrom]**

Community, T. G. (2022b). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.*, *50*(W1), W345–W351. https://doi.

org/10.1093/nar/gkac247 **[cito:usesDataFrom]**

Goble, C., Soiland-Reyes, S., Bacall, F., Owen, S., Williams, A., Eguinoa, I., Droesbeke, B., Leo, S., Pireddu, L., Rodríguez-Navas, L.others. (2021). Implementing FAIR digital objects in the EOSC-life workflow collaboratory. *Zenodo*. **[cito:usesDataFrom]**

Gray, A. J., Goble, C., & Jimenez, R. C. (2017). Bioschemas: From potato salad to protein annotation. *The 16th International Semantic Web Conference 2017*. **[cito:usesMethodIn]**

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. **[cito:usesDataFrom]**

Gustafsson, O., Christiansen, J., Nelson, T., Ward, N., Roberts, D., Davis, B., De La Pierre, M., Chew, T., Gorse, D., Price, G., & Lonie, A. (2021). *Australian BioCommons ToolFinder: discovery of bioinformatics software in an Australian infrastructure context*. Zenodo. https://doi.org/10.5281/zenodo.5587837 **[cito:usesMethodIn]**

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(03), 90–95. **[cito:usesMethodIn]**

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H., Kalaš, M., Schwämmle, V., Grüning, B., Beard, N., Lopez, R., Duvaud, S., Stockinger, H., Persson, B., Vařeková, R. S., Raček, T., Vondrášek, J., Peterson, H., Salumets, A., Jonassen, I., . . . Brunak, S. (2019). The bio.tools registry of software tools and data resources for the life sciences. *Genome Biology*, *20*(1), 164. https://doi.org/10.1186/s13059-019-1772-6 **[cito:usesDataFrom]**

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., & Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/bioinformatics/btt113 **[cito:usesDataFrom]**

Ison, J., Kalaš, M., Ménager, H., Willighagen, E., Grüning, B., & albangaignard. (2020). *Edamontology/edamontology: EDAM 1.25* (Version 1.25) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3899895 **[cito:usesDataFrom]**

Jaaniso, E. (2016). Automatic mapping of free texts to bioinformatics ontology terms. *Master's Thesis (30 ECTS), University of Tartu, Institute of Computer Science, Computer Science Curriculum*. **[cito:describes]**

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & team, J. development. (2016). Jupyter notebooks - a publishing format for reproducible computational workflows. In F. Loizides & B. Scmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (pp. 87–90). IOS Press. https://eprints.soton.ac.uk/403913/ **[cito:usesMethodIn]**

Price, G., & Farquharson, K. (2022). *PacBio HiFi genome assembly using hifiasm v2.1*. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.221.3 **[cito:usesDataFrom]**

Sloggett, C., Goonasekera, N., & Afgan, E. (2013). BioBlend: automating pipeline analyses within Galaxy and CloudMan. *Bioinformatics*, *29*(13), 1685–1686. https://doi.org/10.1093/bioinformatics/btt199 **[cito:usesDataFrom]**

Veiga Leprevost, F. da, Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., & Perez-Riverol, Y. (2017). BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, *33*(16), 2580–2582. https://doi.org/10.1093/bioinformatics/btx192 **[cito:usesDataFrom]**