How to improve the annotation of Galaxy resources? Outcomes of an online hackathon for improving the annotation of Galaxy resources for microbial data resources

Bérénice Batut, Matthias Bernt, Mina Hojat Ansari, Matúš Kalaš, Paul Klemm, Romane Libouban, Engy Nasr, Claire Rioualen, Wai Cheng Thang, Rand Zoabi, et al.

BioHackrXiv

How to improve the annotation of Galaxy resources? Outcomes of an online hackathon for improving the annotation of Galaxy resources for microbial data resources

Bérénice Batut[1, 2], Matthias Bernt[3], Mina Hojat Ansari[4], Matúš Kalaš[5], Paul Klemm[6], Romane Libouban[7], Engy Nasr[4], Claire Rioualen[1], Wai Cheng Thang[8, 9], Rand Zoabi[4], and Paul Zierep[4]

1 Institut Français de Bioinformatique, CNRS UAR 3601, Évry, France **2** Mésocentre Clermont-Auvergne, Université Clermont Auvergne, Aubiere, France **3** Department Computational Biology & Chemistry, UFZ - Helmholtz Centre for Environmental Research, Leipzig, Germany **4** Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany **5** Department of Informatics, University of Bergen, Norway; and ELIXIR Norway **6** Center for Synthetic Microbiology (SYNMIKRO), Philipps-University Marburg, Marburg, Germany **7** GenOuest, University of Rennes, INRIA, CNRS, IRISA, Rennes, France **8** Queensland Cyber Infrastructure Foundation (QCIF), Australia **9** Institute of Molecular Bioscience, University of Queensland, St Lucia, Australia

BioHackathon series:
Improving the annotation of Galaxy resources for microbial data resources
Online
BioHackrXiv

Submitted: 29 Apr 2024

Published by BioHackrXiv.org

Introduction

Galaxy (The Galaxy Community, 2022) is a web-based data analysis platform initially developed for bioinformatics but that expanded its scope over the years to become a global and cross-domain community.

Galaxy offers almost 10,000 different tools. To solve the visibility of tools across this ecosystem, a pipeline (Galaxy Tool Metadata Extractor) (Zierep et al., 2024) was developed at the ELIXIR BioHackathon Europe 2023 that collects Galaxy wrappers from a list of GitHub repositories and automatically extracts their metadata (including Conda version (*Anaconda Software Distribution*, n.d.), bio.tools identifier Ison et al. (2019), BIII identifier (Paul-Gilloteaux et al., 2021), and EDAM ontology (Black et al., 2022), tool availability on public servers, usage statistics on the Europen Galaxy Server). The pipeline can filter its inputs to only include tools that are relevant to a specific research community.

While developing this pipeline, we realized many tools miss proper annotations. Parallel to the pipeline development, annotation guidelines were established to solve this issue. An effort was started to update 50+ Galaxy tools, link them to their respective bio.tools entries, and collectively peer-review the results. However, that was far from enough to properly annotate all Galaxy tools and other types of Galaxy resources like training material and workflows.

microGalaxy is a community of practice within the Galaxy community that focuses on resources related to microbial research. Using Galaxy Tool Metadata Extractor, the microGalaxy community obtained a list of 260+ Galaxy tools related to microbial data analysis. Within this list, only half of them were linked to bio.tools entries, and many others had vague or very generic EDAM annotations in their bio.tools entries. The microGalaxy community decided to continue the work of improving the connection between Galaxy tools and bio.tools and improving the bio.tools entries.

In addition to the tools, the microGalaxy community offers also other resources (30+ tutorials, 30+ workflows) that are also not properly annotated using ontologies like EDAM. Annotating all mentioned resources would improve their findability but also allow for aggregation and display of resources covering similar topics.

To facilitate this work and work on a proof-of-concept for other communities, the microGalaxy community organized an online hackathon to improve the annotation of Galaxy resources and expand Galaxy Tool Metadata Extractor to other resources like tutorials and workflows.

Methods

Initial objectives

The initial objectives of this hackathon were to improve the annotation of the Galaxy resources (tools, training, workflows) for microbial data analysis by:

- Linking microbial Galaxy tools to bio.tools to obtain EDAM ontology annotation
- Improving bio.tools annotations
- Annotating existing microbial-related tutorials with EDAM terms
- Reflecting on the addition of EDAM terms to workflows
- Reflecting on missing terms in the EDAM ontology for microbial data analyses
- Brainstorming on a way to connect tool annotations to improve training and workflow annotations

Organization

The hackathon was held online from March 11th to 15th 2024. It was free of cost. A Zoom room was open the whole week from 9am to 5pm CET. The microGalaxy Matrix chat was also used for communication. Two daily stand-ups were done (one at 9:30 am CET and one at 4:00 pm CET) to accommodate different time zones. Additionally two brainstorming meetings on crucial subjects were planned.

The hackathon was widely advertised by creating an event on the Galaxy Hub website, the IFB catalog and de.NBI website, by sharing the information on different mailing lists and several community channels (Slacks or Matrix), and by spreading on social media.

Coordination

The work during the week was coordinated using a single Google document that contained all information related to this hackathon (schedule, template for advertisement, pre-registration, etc). In this document, we also created a "How to contribute" section with step-by-step information on the different possible ways to contribute, in particular using the tutorial.

In order to track the work and avoid work duplication, a "tracking" spreadsheet was created before the event. In this spreadsheet, several sheets were created and pre-filled: (i) "Galaxy Tools not linked to bio.tools id" with their metadata to track the bio.tools entry creation and the linking between bio.tools and the Galaxy tool, (ii) "Galaxy Tools linked to bio.tools id" to track bio.tools EDAM annotation updates, (iii) "Microbial-related Galaxy Tools" to review the curated list of microbial-related tools, (iv) "Microbial-related Tutorials" to track the EDAM topics identification and tutorial update. A last sheet was added to get an overview of the progress in the different sheets.

This sheet along with the "How to contribute" section was introduced on the first day and to any newcomer during the week.

The daily stand-up meetings were really useful for the coordination. In the 9:30 am CET meetings, participants from the APAC time zone shared their achievements for the day and participants from the EMEA time zone shared their objectives for the day. In the 4:00 pm CET meetings, participants from the EMEA time zone shared their achievements for the day.

Outcomes

Participants

During the week, 15+ people from all over the world joined the effort, either just to discuss and learn about our efforts or more actively. It was a great opportunity to connect with members of other communities like QIIME2 or NFDI4Microbiota.

The most active contributors (10 persons) were from Australia, Germany, France, and Norway.

Annotation of the microbial-related resources

Improving the annotation of the Galaxy resources for microbial data analysis was the main objective of this hackathon. The outcomes are far beyond what we expected.

Linking Galaxy Tools to bio.tools entries

Before this hackathon, 98 Galaxy tools (of the 200 microbial-related tools) were not linked to bio.tools, either because no bio.tools entry exists or because it was not provided in the Galaxy tool itself.

During the hackathon, **41** tools have been added to bio.tools.

Most of the Galaxy tools have been linked to their corresponding bio.tools entries with the actual status being:

Status	Galaxy tools
Pull Request merged	33
Pull Request created, but not merged	53
Tools left	12

The tools left are complicated cases where tools are deprecated or duplicated.

For the other tools, we encountered several issues during the hackathon that delayed the merging of the Pull Request. First, a lot of tools were not following current best practices for Galaxy Tool development so the linting of the continuous integration was failing. Once that was solved, tests were also failing for a bunch of tools. Fixing linting and test errors slowed down the process of the tool annotation but improved the quality of Galaxy tools. Secondly, Galaxy tool sources are stored in GitHub repositories where none of the current contributors have maintainer rights.

Review of bio.tools entries

For the Galaxy tools already linked to bio.tools, we started to review the EDAM annotations in the corresponding bio.tools entries.

Status	bio.tools entries
Reviewed	25
Ongoing review	14
To review	103

Over 140+ Galaxy tools linked to bio.tools entries, we managed to review 25 bio.tools entries. 14 are currently ongoing, especially because we are waiting to have edit access to the bio.tools entries.

For most of the tools already or under review, their EDAM annotations have to be updated because they are too large or missing some functionalities. We then need to continue the work started here to be able to use these annotations for offering a comprehensive list of tools given EDAM Topics and Operations.

Microbial-related Galaxy Tools

Galaxy Tool Metadata Extractor extracted 400+ that might be of interest for microbial data analysis. We did our first curation during the BioHackathon in November 2023. During this hackathon, some fresh pair of eyes started to review the curated list.

Status	Microbial-related Galaxy tool
To keep	180
To exclude	145
To review	420

We will need to continue this work but we might use information from workflows and tutorials to automatically identify the essential tools to keep in this list.

Microbial-related Tutorials

On the Galaxy Training Network (GTN) Materials Hiltemann et al. (2023), the microGalaxy community offers 30+ tutorials. These tutorials were not annotated with EDAM Topic or Operation.

During the hackathon, 33 tutorials were matched with EDAM topics, that were added using Pull Requests to GTN GitHub repository. A similar effort is ongoing to annotate the topics offered on the GTN website with corresponding EDAM Topics.

For EDAM Operation in tutorials, another approach will be implemented. Each tutorial comes with a list of tools. This list will be compared to the information extracted by Galaxy Tool Metadata Extractor and EDAM Operations of the tools will be added to the EDAM Operation of the tutorials.

Improvements to Galaxy Tool Metadata Extractor

Feedback from the annotation process was used to improve the Galaxy Tool Metadata Extractor.

Improvement of the bio.tools id extraction

No bio.tools IDs were extracted when there was unnecessary white space in the Galaxy tool XML. It was fixed in Galaxy Tool Metadata Extractor code (Pull request #1, Pull request #2). In the Galaxy code itself, two new tool linting rules have been added: one to check the validity of bio.tools annotation is Galaxy tools and one to check for leading and trailing spaces in the text of leaf elements in Galaxy tool XML files. These caused problems in the metadata extractor.

Some tool suites like SPAdes include multiple bio.tools references to individual tools. To track those cases all bio.tools references for a tool are collected and stored in an additional column.

New GitHub repository sources

To extract tool suites, the Galaxy Tool Metadata Extractor uses the GitHub repositories which are processed by the planemo monitor. So tools stored in other repositories were not extracted and available in the tool tables. A configuration option was added that allows to add additional repositories, such as the QIIME2 repository (Bolyen et al., 2019).

Availability on all public servers

Initially, tools within a suite were checked if they were installed on the largest Galaxy servers (usegalaxy.eu, usegalaxy.org, usegalaxy.org.au). The ratio of installed tools on these servers was given in the suite row with one column per server.

The approach was slightly modified. The tools within a suite are now checked if installed on the list of all public servers. If at least one of them is installed, the server is listed in the new column `All Server Availability`.

For each UseGalaxy server, a column is added with the number of tools within the suite that is installed on the corresponding server.

Usage statistics

Usage statistics of the tools on UseGalaxy Europe are extracted to obtain the overall tool usage as well as how many users executed the tool in the last two years. However, for some tools, these usage statistics could not be extracted.

This is due to incorrect matching between the name of the wrapper suite and the name of the tool on the server. It has been fixed by matching the name of the individual tool IDs included in the suite with the tool ID on the server. This allowed us to extract the tool usage statistics for all tools that are indeed installed on UseGalaxy Europe.

Tracking of tool extraction

To track the progress in the number of extracted tools and their annotation, a script was developed to collect the tables generated by Galaxy Tool Metadata Extractor from every GitHub commit, including the automatic commits that are created by the bot every week to update the results with the newest tools as well as any commit that was used to improve the code.

The script generates a plot that shows the evolution of time of the number of Galaxy suites, Galaxy suites linked to bio.tools entries, Galaxy suites with usage statistics, Galaxy suites with associated conda package, and the number of the tools installed on the UseGalaxy servers.

Progress Image

Landing page and interactive table update

The landing page and interactive table are also being updated to make use of the ELIXIR Toolkit Theme. That will give a nicer table and page while being sure that the table can still be embedded into other webpages.

Reflexion on EDAM terms for microbiome

In the latest version of EDAM, new topics have been added, including Multiomics (topic_4021) and metabarcoding (topic_4038). But there are still some missing terms to represent what certain tools or training are doing and for what they are useful.

During the hackathon, we had a dedicated meeting to reflect on the EDAM terms for microbiome research. We suggested that some operations could become topics: Genome annotation, Genome assembly (if merged with the Sequence assembly topic), Multilocus sequence typing / Molecular typing / Genotyping.

We also reflected on a new Microbiome or Biome (or maybe both) topic that could encompass several subtopics: Metagenomics, Metabarcoding, Metatranscriptomics, Metaproteomics, etc.

Regarding assembly, "Sequence assembly" exists as a topic and an operation, and "Genome assembly" is an operation. But there are no concepts to represent Metagenomics assembly. It could be added as a new operation or topic.

Finally, we discussed binning. There is a "Read binning" operation mentioning contigs in its description. It could be renamed to "Binning" only to avoid confusion and include two suboperations: taxonomic binning and composition binning.

More microbial related work

During the hackathon, contributions went also beyond resource annotation and Galaxy Tool Metadata Extractor.

DADA2

A workflow for DADA2 (Callahan et al., 2016) has been finalized for IWC, a GitHub repository with Galaxy Workflows maintained by the Intergalactic Workflow Commission, and is now listed in two workflows registries: Dockstore(Yuen et al., 2021) and WorkflowHub(Goble et al., 2021).

A discussion with the DADA2 community has been initialized in how the existing DADA2 tutorial for R may be reused as a tutorial in the Galaxy Training material.

BIOM

A tool has been added to allow the import of BIOM files into phyloseq objects and a bug in the BIOM format tools has been fixed.

QIIME2

A Galaxy metadata setting problem (for an unrelated datatype) has been fixed which made it impossible to upload QIIME2 reference data to Galaxy.

QIIME2 workflows are being added to IWC so they can be listed on WorkflowHub and Dockstore.

A new QIIME2 tutorial has been linked in the Galaxy Training Network. The training materials have also been discussed with the QIIME2 developers. As soon as more QIIME2 tutorials are available for Galaxy, more links will be created in the GTN.

Other points were discussed: how reference data can be integrated into the QIIME2 Galaxy tools, how data parameters can be better annotated in the QIIME2 Galaxy tools, and possibilities for parallelization in upcoming QIIME2 releases.

For the improvement of the datatype usage, a script has been developed that extracts the hierarchy of the Galaxy datatypes and their EDAM data and format annotations which may help the QIIME2 developers to develop the mapping.

Restructuring of the Microbiome Training Topic

The Microbiome Topic that stored all microbiome-related resources of the Galaxy Training Network has been restructured. A better structure was obtained by adding subtopics: Metabarcoding, Metagenomics, and Metatranscriptomics.

Conclusion and outlook

The hackathon was successful with outcomes beyond the initial expectations. This was only possible as a community effort.

Using the work done during this hackathon, we can aggregate microbial-related Galaxy tool suites using their EDAM Topics and Operation and give more visibility to the available tools for different analyses. This can be done analogously for training resources and hopefully later for workflows. The Galaxy Tool Metadata extractor could then go beyond Galaxy tools and generate metadata for communities as a series of tables with curated and annotated Galaxy resources (tools, training, and workflows). These tables could be used in the community pages to give more visibility to the amazing work they are doing and help users find the resources they need.

Acknowledgements

The Institut Français de Bioinformatique (IFB) is funded by the Programme d'Investissements d'Avenir (PIA), grant Agence Nationale de la Recherche, number ANR-11-INBS-0013.

References

Anaconda software distribution. (n.d.). Retrieved November 1, 2016, from https://anaconda.com **[cito:citesAsAuthority]**

Batut, B., Hiltemann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Blank, C., Bretaudeau, A., Brillet-Guéguen, L., Čech, M., Chilton, J., Clements, D., Doppelt-Azeroual, O., Erxleben, A., Freeberg, M. A., Gladman, S., Hoogstrate, Y., Hotz, H.-R., Houwaart, T., Jagtap, P., … Grüning, B. (2018). Community-Driven Data Analysis Training for Biology. *Cell Systems*, *6*(6), 752–758.e1. https://doi.org/10.1016/j.cels.2018.05.012 **[cito:citesAsAuthority]**

Black, M., Lamothe, L., Hager Eldakroury, Kierkegaard, M., Ankita Priya, Machinda, A., Khanduja, U. S., Drashti Patoliya, Rashika Rathi, Tawah Peggy Che Nico, Umutesi, G., Blankenburg, C., Op, A., Chieke, P., Omodolapo Babatunde, Laurie, S., Neumann, S., Schwämmle, V., Kuzmin, I., … Matúš Kalaš. (2022). EDAM: The bioscientific data analysis ontology (update 2021)[version 1; not peer reviewed]. *F1000Research*. https://doi.org/10.7490/f1000research.1118900.1 **[cito:usesDataFrom]**

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F.others. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9 **[cito:citesAsAuthority]**

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869 **[cito:citesAsAuthority]**

Goble, C., Soiland-Reyes, S., Bacall, F., Owen, S., Williams, A., Eguinoa, I., Droesbeke, B., Leo, S., Pireddu, L., Rodríguez-Navas, L., Fernández, J., Capella-Gutierrez, S., Ménager, H., Grüning, B., Serrano-Solano, B., Ewels, P., & Coppens, F. (2021). *Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory*. https://doi.org/10.5281/zenodo.4605654 **[cito:citesAsAuthority]**

Hiltemann, S., Rasche, H., Gladman, S., Hotz, H.-R., Larivière, D., Blankenberg, D., Jagtap, P. D., Wollmann, T., Bretaudeau, A., Goué, N., Griffin, T. J., Royaux, C., Bras, Y. L., Mehta, S., Syme, A., Coppens, F., Droesbeke, B., Soranzo, N., Bacon, W., … and, B. B. (2023). Galaxy Training: A powerful framework for teaching! *PLoS Computational Biology*, *19*(1), e1010752. https://doi.org/10.1371/journal.pcbi.1010752 **[cito:citesAsAuthority]**

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). https://doi.org/10.1186/s13059-019-1772-6 **[cito:usesDataFrom]**

Ison, J., Ienasescu, H., Rydza, E., Chmura, P., Rapacki, K., Gaignard, A., Schwämmle, V., Helden, J. van, Kalaš, M., & Ménager, H. (2021). biotoolsSchema: a formalized schema for bioinformatics software description. *GigaScience*, *10*(1), giaa157. https://doi.org/10.1093/gigascience/giaa157 **[cito:citesAsAuthority]**

Paul-Gilloteaux, P., Tosi, S., Hériché, J.-K., Gaignard, A., Ménager, H., Marée, R., Baecker, V., Klemm, A., Kalaš, M., Zhang, C., Miura, K., & Colombelli, J. (2021). Bioimage analysis workflows: Community resources to navigate through a complex ecosystem. *F1000Research*, *10*, 320. https://doi.org/10.12688/f1000research.52569.1 **[cito:citesAsAuthority]**

The Galaxy Community. (2022). 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:citesAsAuthority]**

Yuen, D., Cabansay, L., Duncan, A., Luu, G., Hogue, G., Overbeck, C., Perez, N., Shands, W., Steinberg, D., Reid, C., Olunwa, N., Hansen, R., Sheets, E., O'Farrell, A., Cullion, K., O'Connor, B. D., Paten, B., & Stein, L. (2021). The Dockstore: Enhancing a community platform for sharing reproducible and accessible computational protocols. *Nucleic Acids Research*, *49*(W1), W624–W632. https://doi.org/10.1093/nar/gkab346 **[cito:citesAsAuthority]**

Zierep, P., Batut, B., Kalaš, M., Kayikcioglu, T., Nasr, E., Soranzo, N., Thang, W. C., Wang, J., & Gustafsson, O. J. R. (2024). *How to increase the findability, visibility, and impact of galaxy tools for your scientific community*. BioHackrXiv. https://doi.org/10.37044/osf.io/qjbxc **[cito:citesAsAuthority]**