How to increase the findability, visibility, and impact of Galaxy tools for your scientific community

Paul Zierep, Bérénice Batut, Matúš Kalaš, Tunc Kayikcioglu, Engy Nasr, Nicola Soranzo, Wai Cheng Thang, Joseph Wang, Ove Johan Ragnar Gustafsson

BioHackrXiv

How to increase the findability, visibility, and impact of Galaxy tools for your scientific community

Paul Zierep[1, a], **Bérénice Batut**[1, 2, a], **Matúš Kalaš**[3], **Tunc Kayikcioglu**[1], **Engy Nasr**[1], **Nicola Soranzo**[4], **Wai Cheng Thang**[5, 6], **Joseph Wang**[7], **and Ove Johan Ragnar Gustafsson**[8]

1 Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany **2** Institut Français de Bioinformatique, CNRS UAR 3601, Évry, France & Mésocentre Clermont-Auvergne, Université Clermont Auvergne, Aubiere, France **3** Department of Informatics, University of Bergen, Norway; and ELIXIR Norway **4** Earlham Institute, Norwich Research Park, Norwich, UK **5** Queensland Cyber Infrastructure Foundation (QCIF), Australia **6** Institute of Molecular Bioscience, University of Queensland, St Lucia, Australia **7** Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, Queensland, Australia **8** Australian BioCommons, University of Melbourne, Melbourne, Victoria, Australia **a** These authors contributed equally to this work

BioHackathon series:
BioHackathon Europe 2023
Barcelona, Spain, 2023
Project 25 - Increasing the findability, visibility, and impact of Galaxy tools for specialised scientific communities

Submitted: 02 Apr 2024

Published by BioHackrXiv.org

Introduction or Background

Galaxy (The Galaxy Community, 2022) is a web-based analysis platform offering almost 10,000 different tools, which are developed in various GitHub repositories.
Furthermore, the Galaxy community embraces granular implementation of software tools as sub-modules. In practice, this means that tool suites are separated into sets of Galaxy tools, also known as Galaxy wrappers, that contain the functionality of a corresponding sub-component. Some key examples of tool suites include QIIME 2 (Bolyen et al., 2019) and OpenMS (Röst et al., 2016), which translate to tens and even hundreds of Galaxy tools. While granularity supports the composability of tools into diverse purpose-specific workflows, this decentralised development and modular architecture can make it difficult for Galaxy users to find and use tools. It may also result in Galaxy tool-wrapper developers duplicating efforts by simultaneously wrapping the same software. This is further complicated by the scarcity of tool metadata, which prevents filtering of tools as relevant for a specific scientific community or domain, and makes it impossible to employ advanced filtering by ontology terms like the ones from EDAM (Black et al., 2022). The final challenge is also an opportunity: the global and cross-domain nature of Galaxy means that it is a big community. Solving the visibility of tools across this "ecosystem", and the resulting benefits, are far-reaching for the global collaborative development of data-analysis tools and workflows.

To provide the scientific community with a comprehensive list of annotated Galaxy tools, we developed a pipeline 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. (2021)](Ison et al., 2019), BIII identifier (Paul-Gilloteaux et al., 2021), and EDAM annotations). The workflow also queries the availability of the tools from the three main Galaxy servers (usegalaxy.*) as well as usage statistics from usegalaxy.eu (Note: the other main Galaxy servers, usegalaxy.org and usegalaxy.org.au, will be queried for usage statistics in coming updates).

Crucially, the pipeline can filter its inputs to only include tools that are relevant to a specific research community. Based on the selected filters, a community-specific interactive table is generated that can be embedded, e.g. into the respective Galaxy Hub webpage or Galaxy subdomain. This table allows further filtering and searching for fine-grained tool selection. The pipeline is fully automated and executes weekly. Any scientific community can apply the pipeline to create a table specific to their needs.

An interactive table that presents metadata is only as useful as the metadata annotations it is capturing. To improve the metadata coverage for the interactive table, the project also directly addressed the quality of tool annotations in bio.tools for the microGalaxy community: a community with a focus on tools related to microbial research.

Annotation guidelines were established for this purpose, the process of updating Galaxy tool wrappers to include bio.tools identifiers was started and the outcome of these activities was evaluated using a crowdsourced approach. During the BioHackathon Europe 2023 week, the annotation practices were applied to the tools selected from the microGalaxy community. This effort allowed the team to connect more than 50 tools to their respective bio.tools entry, update the registry entry, and collectively peer-review the results.

The established pipeline and the annotation guidelines can support any scientific community to make their Galaxy tools more findable, visible, comparable, understandable, and accessible. Here, we describe the methods and processes that resulted from this project and highlight how this will now allow the microGalaxy community to confidently navigate an ever-expanding landscape of research software in the Galaxy framework.

Methods

Domain-specific interactive tools table

To create the domain-specific interactive tools table, Galaxy tool-wrapper suites are first parsed from across multiple GitHub repositories. In effect, the repositories monitored by the planemo-monitor (Bray et al., 2022) are scraped using a custom script. The planemo-monitor is part of the Galaxy tool-update infrastructure and keeps track of the most up-to-date tool development repositories.

Metadata is extracted from each parsed tool-wrapper suite. This includes wrapper suite ID, scientific category, Bioconda dependency, and a repository URL from bio.tools. As a tool suite can be composed of multiple individual tools, the tool IDs for each tool are also extracted. The bio.tools reference is used to request metadata annotations via the bio.tools API, including bio.tools description and functionality annotation using EDAM ontology concepts (Black et al., 2022). The latest Conda package version is retrieved via the Bioconda API and compared to the Galaxy tool version to determine the tool's update state (i.e. to update, or no update required).

The Galaxy API is used to query if each tool is installed on one of the three UseGalaxy servers (usegalaxy.eu, usegalaxy.org, usegalaxy.org.au). Furthermore, the tool usage statistics can be retrieved from an SQL query that needs to be executed by Galaxy administrators. The query used in the current implementation shows the overall tool usage as well as how many users executed a tool in the last 2 years on the European server (usegalaxy.eu).

The output of the pipeline is a table that combines Galaxy wrappers with their metadata. The complete table can then be filtered to include only tools with relevance for specific communities. The initial filtering step is based on the scientific category, which is defined for every Galaxy tool wrapper. These categories are high-level and cannot distinguish between specific tool functions. However, they allow for the isolation of a subgroup of the initial table for further curation. The filtered table can then be manually curated by community curators. This curation step involves annotating which of the extracted tools should be kept in the final table. Curators can use the EDAM annotations and tool descriptions to assist with this curation step. The `to keep` labels for each tool are stored to reduce the replication of effort even further. The practical outcome is that for repeat executions of the workflow, only new tools require curation.

The curated tools are transformed into an interactive web table using the data tables framework(*DataTables | Table Plug-in for jQuery*, n.d.). The table is hosted on GitHub and deployed via GitHub pages for each community. This implementation enables complex queries and

filtering without the need for a database backend. The table can be embedded in any website via an iframe: examples include the Galaxy community Hub page for microGalaxy or the microGalaxy subdomain. Furthermore, a word cloud based on the usage statistics of the tools is created.

The workflow is run weekly via GitHub Actions continuous integration, providing an up-to-date table for each community. The usage of an iframe enables updates for the table to propagate automatically to any website where it is deployed.

Any Galaxy community can use the pipeline by adding a folder in the project GitHub repository. To initialise the pipeline for a new community you need to add a new subfolder of `data/communities/` , and inside it add a file called 'categories' with a list of Galaxy ToolShed (Blankenberg et al., 2014) categories. Additionally, tools that should be excluded or included after filtering, can be added to respective files as well. A working example of the community configuration files can be found in the folder for the microGalaxy community.

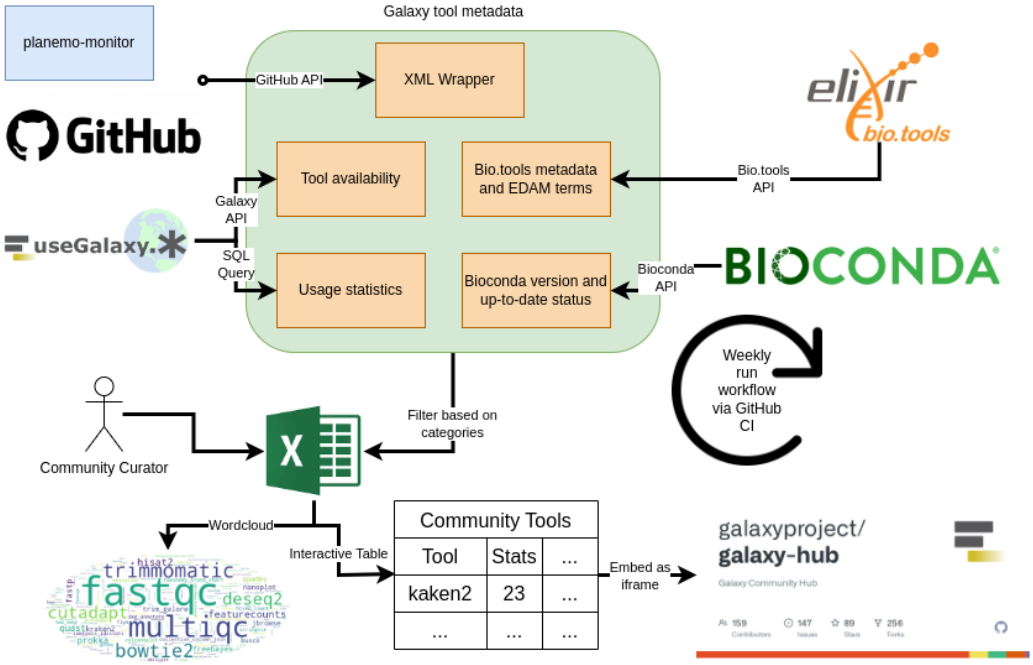

Figure 1: Workflow of the Galaxy tool metadata extractor pipeline. Tool wrappers are parsed from different repositories and additional metadata is retrieved from bio.tools, Bioconda, and the main public Galaxy servers. Upon filtering and manual curation of the data for specific scientific communities, the data is transformed into interactive web tables and a tool usage statistics-based word cloud, that can be integrated into any website.

Annotation workflow

The annotation process begins by selecting a tool from a Galaxy community. This step can make use of the interactive table created by the Galaxy tool extractor scripts presented above. A curator then needs to visit the development repository of the Galaxy tool wrapper and search the XML file for a bio.tools xref snippet (Figure 2).

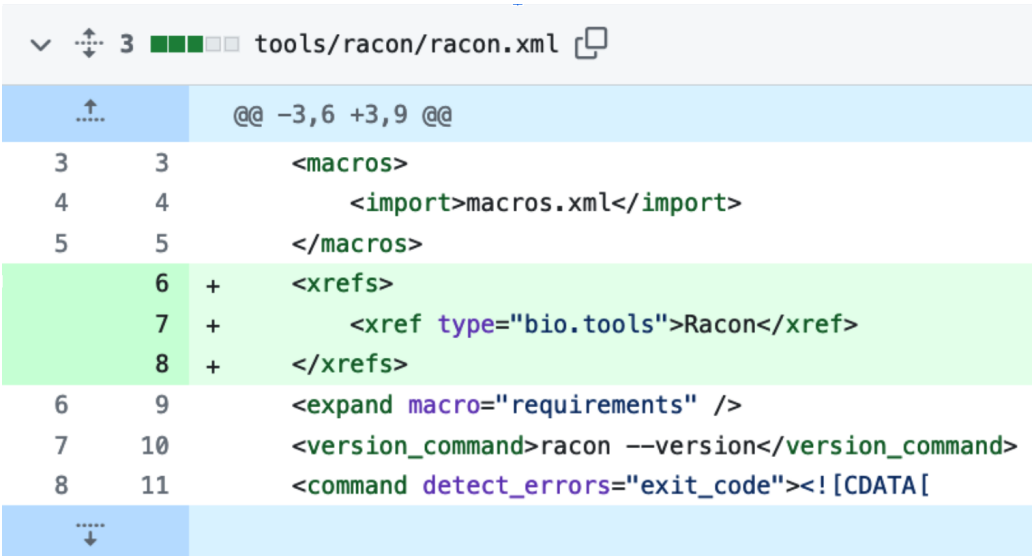

Figure 2: xref snippet example for a Galaxy tool wrapper that contains the tool Racon.

bio.tools is then checked to confirm that a bio.tools identifier does, or does not, exist. The reason for this is that even if a bio.tools identifier exists in a tool wrapper, it may not necessarily exist in bio.tools. This is an observation based on real-world annotation errors and serves as a useful supporting step to improve Galaxy wrapper annotations and the completeness of the bio.tools registry. In addition, if a bio.tools identifier is not included in the wrapper, this does not mean that there is not a bio.tools identifier available in the registry.

There are then two curation paths to choose from, depending on whether a bio.tools identifier exists in the XML wrapper. In both cases, if no bio.tools entry exists, a new entry should be created and updated using the bio.tools wizard. The creation and update of an entry includes adding concepts from the EDAM ontology. This annotation process can be simplified through the use of EDAM Browser [Brancotte et al. (2018)](Eldakroury et al., 2021).

In the case where no bio.tools identifier exists in the Galaxy XML wrapper, the development repository needs to be forked and a new branch created. A new xref snippet can then be added, and a pull request opened against the original repository.

Figure 3 shows a step-by-step breakdown of the above process.

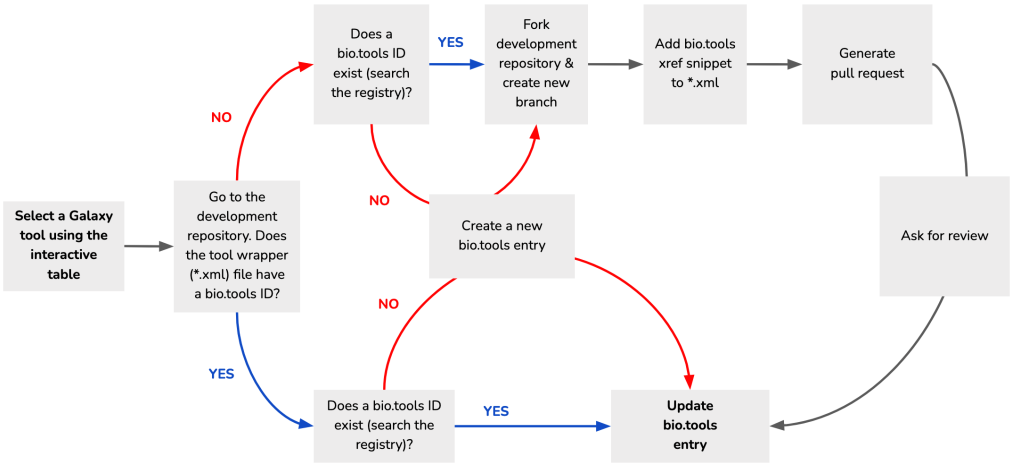

Figure 3: Step-by-step workflow for systematically improving metadata annotations across bio.tools registry entries and Galaxy tool wrappers. After selecting a Galaxy tool and checking for the presence of a bio.tools ID in its XML file, a curator needs to review bio.tools, create a new bio.tools entry (if needed), and then ensure that both this entry and the Galaxy tool XML file are up-to-date. Updating bio.tools makes use of the registry wizard, and updating a Galaxy tool wrapper to include a bio.tools xref snippet requires a pull request against the development repository.

Outcomes and results

There were multiple concrete outcomes from this BioHackathon project, including the ability to create interactive Galaxy tools tables as needed for scientific communities, a process for updating bio.tools, in-development Galaxy Training Network (GTN) tutorials (Batut et al., 2018) describing this process, and an update to the Galaxy IUC tool wrapping standards. These are described in more detail below.

Prototype interactive table for Galaxy communities

The described workflow for the Galaxy tool metadata extractor (see Figure 1) was successfully implemented (GitHub repository) and could extract more than 1,300 Galaxy tool suites (see the GitHub repository pages for an up-to-date table). Of those tool suites, only 267 had a bio.tools identifier, which highlights the importance of performing the annotation process in parallel and complementing the tools with additional metadata. An example view of the created interactive table is shown in Figure 4. As mentioned above, the filtered table for the microGalaxy community has already been embedded in the Hub page for microGalaxy, as well as the dedicated microGalaxy subdomain. The process for creating a new interactive table for a community is currently being transformed into a GTN tutorial.

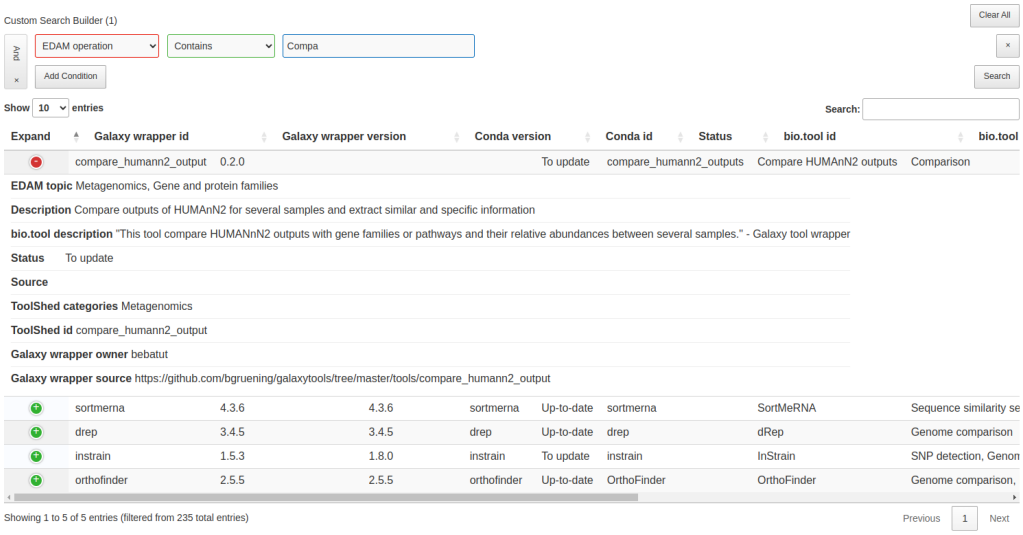

Figure 4: Screenshot of the interactive web table. The table provides comprehensive metadata for all Galaxy wrappers of a specific community and allows for custom searches based on logic filters over all columns. In the shown example, the user queries for all up-to-date tools that are annotated with an EDAM operation that includes "assembly".

bio.tools and EDAM annotations for microGalaxy community

During the week of the BioHackathon, the microGalaxy community executed the annotation workflow as described in the Methods section (see also Figure 3). The initial filtered tool table of the microGalaxy included 218 tool suites, of which 61 had corresponding bio.tools identifiers. The progress of the work was tracked using a GitHub project board. After the annotation process, the number of tools with bio.tools annotations was increased to 107. The added annotations for each respective bio.tools entry were also collectively reviewed by the team. A rerun of the Galaxy tool metadata extractor pipeline collected the additional information, and the metadata is now included in the interactive microGalaxy tool table.

Training materials and updates to standards

To provide the Galaxy research communities with simple and straightforward guide to annotating their respective tool stacks, the described work has been converted into two GTN tutorials:

- Adding and updating best practice metadata for Galaxy tools using the bio.tools registry (Batut et al., 2024)
- Creation of an interactive Galaxy tools table for your community (Batut, 2024)

The guidelines created were also used to update the best practices for creating Galaxy tools of the IUC repository.

Conclusion and outlook

The project was able to successfully meet its aim of creating reusable prototypes and processes that make the richness of the Galaxy tools ecosystem more discoverable and understandable. Central to this work was the Galaxy tool metadata extractor pipeline, which is currently generating comprehensive and interactive tabular summaries of Galaxy tools for the microbial data and image analysis communities within Galaxy (with EU BioHackathon 2023 Project 16). The metadata extractor can be reused by any Galaxy group or community. For example, the biodiversity and ecology community will employ this pipeline in the near future (Waterhouse et al., 2023). The generated tabular tool summary provides valuable information that extends

beyond the use case of listing community tools. Therefore, an integration with the Research Software Ecosystem (RSEc) (Ienasescu et al., 2023) is currently being worked on. Various updates of the Galaxy tool metadata extractor pipeline are also envisioned, such as the integration of comprehensive usage statistics for all large Galaxy servers, additional bio.tools metadata, and a user-friendly integration of manual curation steps.

A set of updates to standards and processes was also created. These will support the ongoing growth of the metadata hosted by the interactive tables: primarily by helping communities to maintain and extend the annotations of Galaxy tool wrappers, and the bio.tools ecosystem on which these wrapper annotations depend.

Acknowledgements

This work was developed as part of BioHackathon Europe 2023. This work was 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.

References

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

Batut, B. (2024). *Creation of an interactive Galaxy tools table for your community (Galaxy Training Materials)*. https://training.galaxyproject.org/training-material/topics/dev/tutorials/community-tool-table/tutorial.html **[cito:citesAsAuthority]**

Batut, B., Gustafsson, J., & Zierep, P. (2024). *Adding and updating best practice metadata for Galaxy tools using the bio.tools registry (Galaxy Training Materials)*. https://training.galaxyproject.org/training-material/topics/dev/tutorials/tool-annotation/tutorial.html **[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:citesAsAuthority] [cito:usesDataFrom]**

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., Team, G., Taylor, J., & Nekrutenko, A. (2014). Dissemination of scientific software with Galaxy ToolShed. *Genome Biology*, *15*, 1–3. **[cito:citesAsAuthority]**

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., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (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]**

Brancotte, B., Blanchet, C., & Ménager, H. (2018). A reusable tree-based web-visualization to browse EDAM ontology, and contribute to it. *Journal of Open Source Software*, *3*(27), 698.

https://doi.org/10.21105/joss.00698 **[cito:citesAsAuthority]**

Bray, S., Bernt, M., Soranzo, N., Beek, M. van den, Batut, B., Rasche, H., Čech, M., Cock, P., Nekrutenko, A., Grüning, B., & Chilton, J. (2022). Planemo: A command-line toolkit for developing, deploying, and executing scientific data analyses. *bioRxiv*. https://doi.org/10.1101/2022.03.13.483965 **[cito:citesAsAuthority]**

DataTables | Table plug-in for jQuery. (n.d.). Retrieved November 28, 2023, from https://datatables.net/

Eldakroury, H., Dhamija, S., Rathi, R., Patoliya, D., Nkwuda, S. C., Singh, G., Yadav, P., D'oleo, K., Cherop, M., Che Nico, T. P., Kalaš, M., Ménager, H., & Brancotte, B. (2021). *EDAM Browser 2.0.0: Browsing multiple versions of EDAM*. Zenodo. https://doi.org/10.5281/zenodo.5808818 **[cito:citesAsAuthority]**

Ienasescu, H., Capella-Gutiérrez, S., Coppens, F., Fernández, J. M., Gaignard, A., Goble, C., Grüning, B., Gustafsson, J., Gelpi, J. L., Harrow, J., Manos, S., Miura, K., Möller, S., Owen, S., Paul-Gilloteaux, P., Peterson, H., Pitoulias, M., Tedds, J., Repchevsky, D., ... Ménager, H. (2023). *The ELIXIR research software ecosystem: An open software metadata commons (BOSC track) [version 1; not peer reviewed]*. F1000 Research. https://doi.org/10.7490/f1000research.1119604.1 **[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]**

Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., ... Kohlbacher, O. (2016). OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, *13*(9), 741–748. https://doi.org/10.1038/nmeth.3959 **[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]**

Waterhouse, R., Adam-Blondon, A., Balech, B., Barta, E., Heil, K., Hughes, G., Jermiin, L., Kalaš, M., Lanfear, J., Pafilis, E., Papageorgiou, A., Psomopoulos, F., Raes, N., Burgin, J., & Gabaldón, T. (2023). The ELIXIR Biodiversity Community: Understanding short- and long-term changes in biodiversity [version 1; peer review: 1 approved with reservations, 1 not approved]. *F1000Research*, *12*(499). https://doi.org/10.12688/f1000research.133724.1 **[cito:citesAsAuthority]**