Using Glittr.org to find, compare and re-use online training materials

Using Glittr.org to find, compare and re-use online training materials

Geert van Geest [1,2*], Yann Haefliger [1], Monique Zahn-Zabal [1], Patricia M. Palagi [1]

[1]Swiss Institute of Bioinformatics, Quartier Sorge - Bâtiment Amphipôle, 1015 Lausanne, Switzerland.

[2]Interfaculty Bioinformatics Unit, University of Bern, Baltzerstrasse 6, 3012, Bern, Switzerland.

*Corresponding author. E-mail: geert.vangeest@sib.swiss

Abstract

A wealth of excellent training materials on computational life sciences is scattered around the internet, but they can be hard to find. Many materials reside in public git repositories that are hosted on platforms such as GitHub and GitLab. Glittr.org aggregates and indexes such git repositories enabling users to find training materials that would otherwise be hard to identify. With the application, users can search and compare training materials based on topic and author, but also on engagement metrics like stargazers (bookmarks) and recency (days since last commit). Glittr.org currently contains 568 entries, which are assigned to six different categories within the field of computational life sciences. By analysing the database, we reveal insights in the availability of materials per topic, collaboration patterns of developers, and licensing practices. This knowledge helps to understand in which areas open training materials are lacking, the importance of git for collaboration on training materials and how licensing can be improved to enhance sharing and reuse. Taken together, we show that Glittr.org contains a wealth of connected and openly available metadata. Therefore, it enhances adherence to the FAIR (Findable, Accessible, Interoperable, Reusable) principles, which benefits the entire life sciences community.

Introduction

The life sciences domain is constantly growing and evolving, requiring professionals to be trained to stay up to date with the most relevant skills and knowledge. To facilitate this, various organizations deliver courses on a large range of topics in many different formats. Many course providers dedicate themselves to actively share their materials with the public, resulting in a growing wealth of public training materials available across the web. This greatly enhances the possibility to re-use these training materials, which is invaluable for both learners and lesson developers; learners can use the material on their own time to obtain required skills, while

lesson developers can use the existing material to build their courses further upon existing ones.

To help developers make their shared material easier to find and re-use, the FAIR principles [1–5] provide guidelines on organizing their materials and metadata. The way training material metadata are represented depends on the platform where they are hosted. Currently, training materials are hosted in diverse locations, like organizational websites, cloud storage, learning management systems, and Git platforms (e.g., GitHub, GitLab). Therefore, metadata representation often does not conform to standards, making it difficult to parse and utilize it effectively. This issue can be relieved by aggregation and indexation, which encourages metadata standardization and therefore enhances findability, re-usability, and integration of metadata over multiple platforms.

Registries that currently aggregate and index training materials in the life sciences include the Open Educational Resources (OER) Commons [6], the training portal of the Global Organisation for Bioinformatics Learning, Education and Training (GOBLET) [7], and ELIXIR's online training registry Training eSupport System (TeSS) [8]. These registries can accommodate any digital object related to education that are accessible through an URL, making them potentially comprehensive. However, this flexibility also comes with disadvantages. Since not every webpage contains standardized metadata, it requires investments from the training material developer to manually enter metadata at the registry or to automate metadata provision, for example by implementing a standardized metadata markup like Bioschemas [9,10].

Using public git repositories for developing and hosting training materials is gaining popularity, especially for materials related to (bio)informatics and data science. It is generally recommended for reproducibility [11] and adopted by many institutions, communities, and individuals (e.g. the Carpentries [12], Galaxy training network [13], the National Bioinformatics Infrastructure Sweden (NBIS) and the SIB Swiss Institute of Bioinformatics. Having training

materials in public Git repositories has major advantages. First, using Git encourages community-driven development. Git is version-control software, and as such, can be used to track changes in files and associate these changes to individuals with distinct levels of permissions. This makes it ideal for co-development and re-use of materials. Using git drives community engagement by supporting direct feedback, suggestions, and corrections. Secondly, the infrastructure around the most popular platform for public git repositories, GitHub (https://github.com) and GitLab (https://about.gitlab.com), are standardized and commonly used. Many people have the know-how to use version control, so the threshold to use or learn how to use git for training material is low – especially in communities related to the informatics field. Thirdly, public Git repositories offer standardized retrieval of a wealth of metadata. Authorship, activity logs, and popularity metrics are all recorded and made accessible through a single point of entry.

Despite many advantages, it remains difficult to find training materials on public git repositories. Although both GitHub and GitLab have sophisticated search functionality, it remains difficult to identify and compare repositories with training materials on a topic of interest. The findability of repositories with training materials depends on the effort of the author to describe the repository, for example in the repository description, the name, the associated website, and repository tags. Since not all authors annotate their repositories extensively and the way developers annotate repositories is not standardized, many repositories with excellent training materials remain hard to find.

This article introduces Glittr.org, a novel web-based resource designed to enhance the discovery and comparison of reusable bioinformatics training materials hosted on public Git repositories. Addressing the needs of both students and trainers, Glittr.org contains intuitive search functionalities based on pre-defined topics and repository metadata. This approach not only facilitates the users' search based on specific topics but also the refinement of their

search based on factors like the training material's popularity and recency. Moreover, Glittr.org supports the adherence to the FAIR principles by making all repository metadata readily accessible through REST API endpoints, fostering findability and reusability of bioinformatics training materials within the wider research community.

Materials and methods

Data collection and curation

Glittr.org is a manually curated list of bioinformatics training materials from anywhere in the world and in any language. The list items are public GitHub or GitLab repositories. The teaching aim of the material should be a general topic in bioinformatics, not tool usage – meaning that inclusion of tutorials on using single software are discouraged. The reason for this is to ensure Glittr.org provides an overview of metadata on training materials, and not on metadata related to software. The quality of the training materials is not evaluated. Repositories that are potentially eligible to be part of Glittr.org are identified from existing collections, internet searches, specific searches on GitHub, (social media) promotions and user contributions through a web form.

Repository metadata are retrieved from GitHub and GitLab. Among those are the repository name, URL, repository description, author and contributor names, their profile and web site, the number of stargazers (bookmarks), the date of the last push, and the license. These are all programmatically accessible from the REST APIs of GitHub and GitLab.

To enable categorization and improve search functionality in Glittr.org, the training content in the repositories are described with tags. These tags are based on existing ontologies as much as possible. Using ontologies greatly enhances standardization and therefore knowledge sharing. Training material in bioinformatics can cover a very wide range of topics, and using a single ontology to describe them can be problematic. Glittr.org therefore uses multiple

ontologies, aiming to give an as rich as possible topic description for each repository. Currently, the following ontologies are used: EDAM [14], Data Science Education Ontology (http://www.pagestudy.org/DSEO/), Experimental Factor Ontology [15] and FAIR Data Train Ontology (https://bioportal.bioontology.org/ontologies/FDT-O). If a term is considered relevant, but not in any of these ontologies, a term will be used from any other ontology on BioPortal [16], a software on bio.tools [17] or an item at Wikidata [18]. By using multiple ontologies, we have the flexibility to annotate the wide range of training material, but still have good level of standardization in relation to the relevant field.

Each repository is associated with one or more tags by a curator, where the order of tags matters, meaning that the tag that describes the content best is first (the 'main tag') and is then followed by other relevant tags all with equal importance. For example, for a single cell transcriptomics course in R, the tag order could be 'Single-cell sequencing', 'R', 'transcriptomics' and 'RNA-seq'. The first and therefore main tag would be 'Single-cell sequencing', which is used to categorize the material, while the others have equal importance and are used to further describe the material. Each tag falls within a broader category, which are maintained within Glittr.org. Currently, these are: (1) Scripting and languages, (2) Computational methods and pipelines, (3) Omics analysis, (4) Reproducibility and data management, (5) Statistics and machine learning, and (6) Others. The main tag of the above-mentioned single cell course is part of the category 'Omics analysis' and so the course material falls within this category.

Web application

The Glittr.org web application is created using the Laravel (https://laravel.com/) framework and is written in PHP. Repository metadata is actively updated twice daily using Laravel jobs. This approach enables handling of potential issues by implementing delayed retries if the GitHub/GitLab APIs are temporarily unavailable. Each metadata update for a single repository

constitutes a dedicated Laravel job, with the overall management of all repository metadata updates performed with the Laravel Haystack package (https://docs.laravel-haystack.dev/). This package ensures job validation and sends email notifications in case of errors.

All the reactivity supporting list functionalities such as filtering, pagination, and sorting is based on the Livewire (https://livewire.laravel.com/) framework. The JSON API endpoints for repositories and Bioschemas are generated with the Laravel-query-builder (https://spatie.be/docs/laravel-query-builder/v5/introduction) package developed by Spatie (https://spatie.be/open-source). This ensures the query parameters and output results follow the JSON API specification as closely as possible. The backend to manage data (topics and categories, repositories, submissions) is created with the package Jetstream (https://jetstream.laravel.com/introduction.html).

All source code is accessible at https://github.com/sib-swiss/glittr and the website is available through https://glittr.org. The application is a SIB Training resource and is hosted at servers from SIB. As an institute with a long history of working with database life cycles, we are committed to sustaining glittr.org for the time it is serving the community by aligning to the TRUST principles [19].

A snapshot of the front-end website is found in Fig 1. The user can easily access both the repository with course material and the course web site (when applicable). In addition, a link to public information about the author or organization that created the material is provided to retrieve more author information. The topics (tags) covered by the material allows users to quickly find course material relevant to their search. The number of stargazers estimates the popularity of the repository, while the number of days since the last push provides an indication of the recency of the material in the repository. The license which applies to the repository contents indicates to trainers wishing to re-use the material what conditions apply.

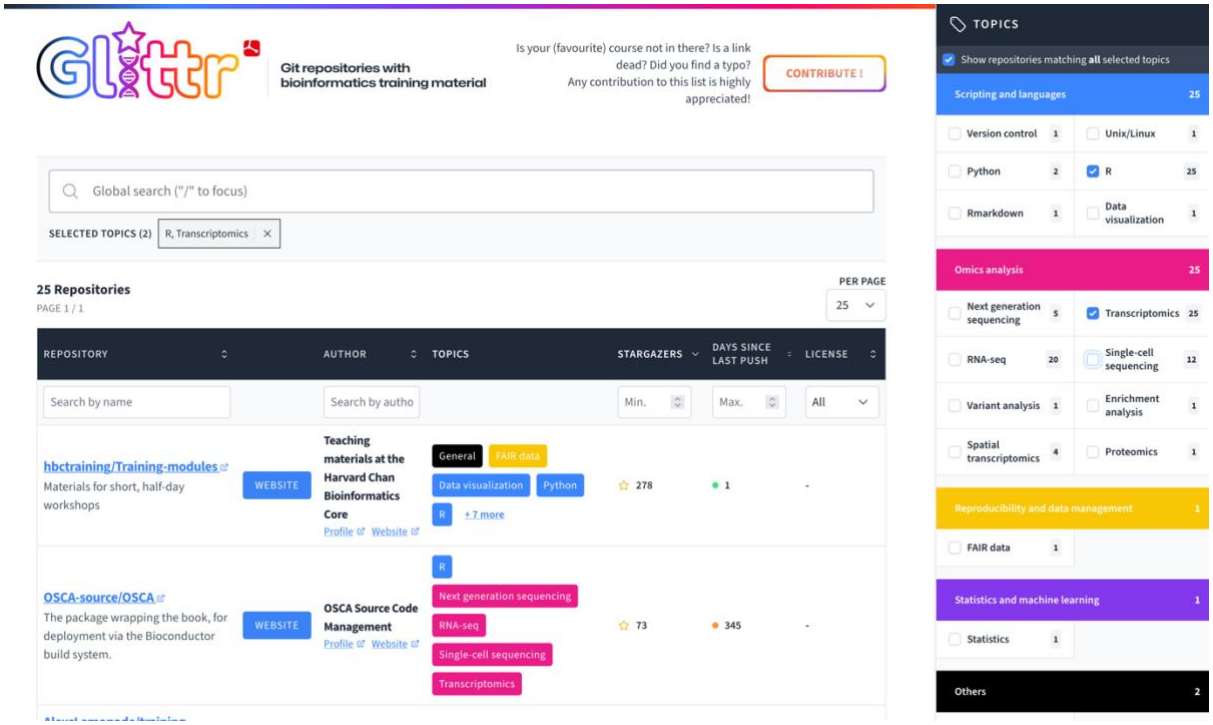

Fig 1. Snapshot of Glittr.org. Here, two tags are selected 'Transcriptomics' and 'R', showing the (current) top 2 of 25 repositories with both tags. By default, results are sorted based on number of 'stargazers', i.e. bookmarks.

The application currently has three REST API endpoints in which data stored by Glittr.org can be programmatically accessed: api/repositories, api/tags and api/bioschemas. The api/repositories endpoint contains all repositories including their metadata as presented on the website. The repositories/tags endpoint contains tag metadata organized by their corresponding category. Tag metadata consists of the ontology, ontology class and description. The api/bioschemas endpoint contains repository and tag metadata according to the Bioschemas [9] TrainingMaterial profile [10] (https://bioschemas.org/profiles/TrainingMaterial).

Training material characteristics

To give an overview and statistics of all the metadata stored on Glittr.org, we used the R httr2 package to access the Glittr.org REST API (see above). Public information on authors, and number of contributors was extracted from the GitHub REST API. Location of authors and

organization was extracted from the 'location' field within the users API. This string was translated in a country with the R package geocode, which has wrappers for the Google Geocoding API. The source code for acquiring the statistics can be found at https://github.com/sib-swiss/glittr-stats and the report can be approached through https://sib-swiss.github.io/glittr-stats/ (updated weekly).

Results and Discussion

There are currently 568 repositories listed on Glittr.org. By combining data from our REST APIs and those of GitHub and GitLab we can give a detailed insight in these repositories. In the first part of this section, we show what kind of repositories with training material are currently available, for which topics training materials exist, what licenses are used, and by whom the repositories are created. In the second part we present several use-cases on how different types of users can interact with Glittr.org to show how trainers, learners and developers can use different functionalities of glittr.org.

Training material topics

Repositories on glittr.org are categorized in six categories and tagged with 55 different topics. All are connected to bioinformatics, but many are not specific to bioinformatics: for example, computer languages are relevant to informatics in general, and statistics to all fields related to science. Therefore, using glittr.org is relevant to a broader public than just persons working in the field of bioinformatics. In this section we give a brief overview of popular categories and topics.

By far most repositories currently on Glittr.org have 'Scripting and languages' as main category – 315 in total (Fig 2A). Within this category, R [20] is the most frequent tag with 230 repositories, followed by Python with 81 repositories (Fig 2C). This suggests that the R language is currently the most popular language to teach and therefore probably to learn in bioinformatics, with

Python coming in second. The popularity of R might not be the only reason; the ease with which one can host static websites within the R framework, with e.g. bookdown [21] or Quarto (https://quarto.org/), could also explain this.

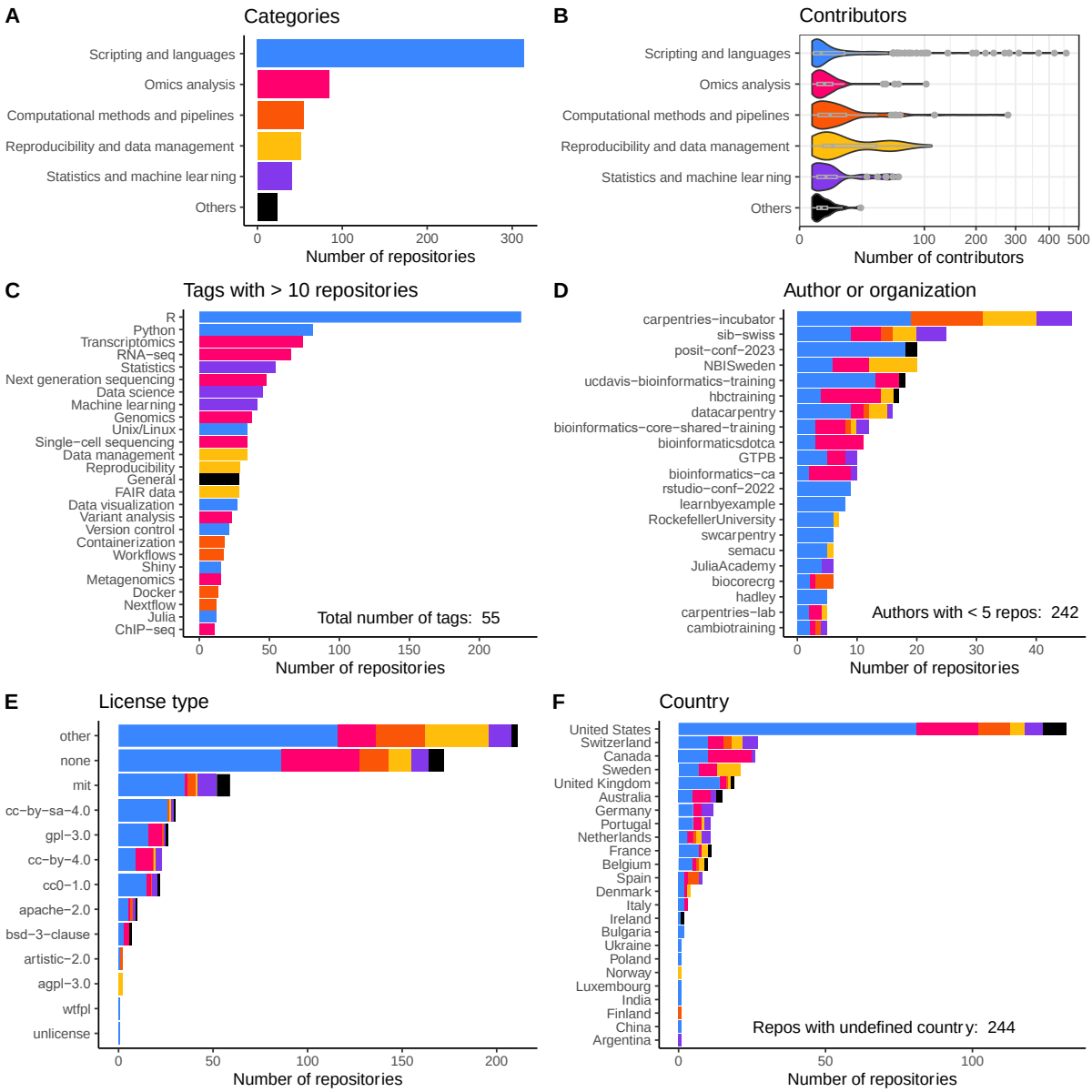

Fig 2. Key statistics by category. Overview of number of repositories by category (A), distribution of number of contributors per repository by category (B), number of repositories per tag, square root transformed (C), per author or organization (D) and per license (E). Colours are according to category as depicted in (A). In C) only the tags with more than 10 repositories are displayed and in D) only the authors/organizations with five or more repositories. At time of writing Glittr.org contained repositories

authored by 242 authors/organizations with less than five repositories. For A), D) and E) the plots are coloured according to the category of the main (i.e. first) tag.

The category 'Omics analysis' is specific to the field of bioinformatics, and many repositories fall within this category. 'Transcriptomics' is the most frequent tag (74 repositories), followed by 'RNA-seq', suggesting an important place for transcriptomics-related courses in current bioinformatics curricula. Transcriptomics analysis skills are popular to acquire, likely because it is one of the most used applications of next generation sequencing, it is applied in many fields in biology [22] and each experiment requires its own analysis choices. Its demand is well served by the vast amount of material available on transcriptomics.

There are 34 repositories on 'Data management', representing its growing importance within the scientific community regarding trends in open and FAIR data. Using public git repositories to develop training materials on data management related topics, is likely to be of particular interest because using git supports transparency and community-based development. Efforts that support open and reproducible science like the ELIXIR Research Data Management (RDM) community, European Open Science Cloud (EOSC) and other research infrastructures likely influence this trend towards using git repositories for training materials both on data management and other topics.

The number of repositories and their diversity show that there is a wealth of public repositories with bioinformatics training material. For many topics, there are high-quality materials in different formats, giving lesson developers and learners many possibilities to re-use existing training materials. Glittr.org organizes this wealth of materials, making them better findable and therefore available for re-use.

Collaboration

An important advantage of using version control for training materials is that it supports collaboration and building further on existing material. This is also represented in the number of collaborators per repository that are on Glittr.org. Most repositories (78.8%) have more than one contributor, meaning the repository is developed by more people than the original author. Many repositories have only a few contributors, between 1 and 5 (59.9%). However, quite a large fraction (25.6%) has more than 10 contributors, showing that these repositories gain from a strong community involvement and therefore increased (re-)usage and quality.

Repositories in the category 'Reproducibility and data management' and 'Computational methods and pipelines' have the largest fraction of repositories developed by a large community (> 50 contributors; Fig 2B). Repositories with a large contributor base within these categories are e.g. elixir-europe/rdmkit - which is a community-created website with RDM best practices [23], galaxyproject/training-material – the Galaxy training network repository with all training materials related to Galaxy [13] - or are forked from a repository with a large community base (e.g. a template from the carpentries [12]). This shows that large and important initiatives rely on git for the development of training materials and that these thrive by cooperation.

Authors and organizations

The origin of repositories on Glittr.org is very diverse; currently, the 568 repositories originate from 263 different GitHub/GitLab authors or organizations. The organization with most repositories is the carpentries-incubator [12], followed by two national institutes, NBIS and SIB (Fig 2C). Obviously, the distribution of repository category varies per organization/author. Bioinformatics institutes, like bioinformatics.ca, Instituto Gulbenkian de Ciência (GTPB), CRUK CI (bioinformatics-core-shared-training), Harvard bioinformatics core (hbctraining), UC Davis Bioinformatics Core (ucdavis-bioinformatics-training), SIB and NBIS, are more focused on

computational biology and have more courses in the category ‘Omics analysis’. Other popular namespaces like carpentries incubator [12], the ’23 posit conference (posit-conf-2023) and learn by example (learnbyexample) are focused on computer skills and therefore have mostly repositories on ‘Scripting and languages’ and ‘Computational methods and pipelines’ (Fig 2C).

In addition to the number of authors, the number of countries the authors affiliate themselves with is also diverse. Current repositories on Glittr.org originate from 24 different countries. Most repositories with the main author or organization having a country defined in its profile originate from the United States, followed by Switzerland and Canada (Fig 2F). This shows that most training materials originate from the Western and particularly English-speaking parts of the world. English is currently the ‘lingua franca’ of the academic world [24], and most training materials in bioinformatics reflect that. Connections between the academic ecosystem and training materials not written in English can be literally lost in translation, and therefore hard to identify. However, there are many communities that demand non-English bioinformatics training material, and we are open to serving these communities by listing them on Glittr.org. Organization of the curation of repositories in other languages is more challenging because it relies on reading and understanding the content, but we encourage initiatives in this direction.

Licensing

Although licensing is an important way to enable and encourage re-use, licensing of public training materials on GitHub and GitLab is not very standardized. Fig 2B gives an overview of licenses per repository that could be recognized as such. For many repositories (172; 30.4%), there is no license detected, and for 211 (37.3%) repositories the license could not be automatically recognized. The most popular recognized license is the MIT license, a software license. The copyright licenses based on Creative Commons (CC) are also quite popular, 75 repositories have a license recognized as such. Since training materials generally consists of text, presentations, and exercises, and not software, it is usually more appropriate to use a

copyright license instead of a software license [4,11]. It therefore seems that many developers of training materials might want to (re-)consider their license.

Use cases

Glittr.org has been developed for use by bioinformatics students, trainers, lesson developers and database developers. To showcase different applications of the resource, we present five different use cases that we have envisaged for Glittr.org. The first four use cases illustrate the ease of use, as only a single click is necessary to obtain an answer, while the last one shows how to retrieve information programmatically.

A student looking for training materials

Our first use case is that of a student looking for training materials, for instance someone wishing to learn the programming language Julia. This high-level, general-purpose dynamic programming language used for numerical analysis and computational science is usually offered in university and higher order education curricula. Using the 'Filter by topic' option to select Julia under the 'Scripting and Languages' section, only repositories which have been manually assigned the topic are retrieved. The top result is the most popular as, by default, the results are displayed in order of decreasing number of stargazers. It is also possible to sort by increasing number of days since the last push to find material which has been recently updated. The link to the website takes the student directly to the page showing the different options to get started with Julia. Glittr.org thus allows students to quickly access the most popular, public course material on the subject.

A trainer looking for material for re-use

This use case is that of a trainer who will be teaching a course on Snakemake and is looking for materials that can be re-used. Snakemake is a workflow management system to create reproducible and scalable data analyses. Selecting Snakemake in the Computational methods

and pipelines section of the 'Filter by topic' option retrieves several repositories with materials to teach Snakemake. By looking at the license column, the trainer can easily spot that there are two repositories that contain material that have a re-use permissive license, one under the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0) and another under the GNU General Public License version 3 license (pl-3.0). For other repositories in the list, the license could not be automatically identified, and therefore the license has been indicated as 'other'. For these repositories some further investigation is needed before they can be re-used. The trainer has quickly found re-usable material on the topic of interest using the Glittr.org web site.

A trainer sharing their own materials

The third use case is that of a trainer who has developed a course on a hot topic in bioinformatics on GitHub or GitLab and wishes to share the material. Contributions to the Glittr.org web site from the training community are welcome and encouraged. The Contribute button on the top right of the home page (Fig 1) links to a simple form. By completing the submission form, the trainer's repository will be added (or updated) to the collection provided it meets the inclusion criteria (public GitHub or GitLab repositories containing bioinformatics training material). The trainer's material is now findable by being included in the Glittr.org web site. Obviously, this does not prevent the trainer from posting the material in other repositories which is, in fact, recommended by the FAIR principles (the more the material is registered, more chances are that they are findable).

A database developer wanting to programmatically access Glittr.org

A final use case is that of a developer who must programmatically access one or more subsets of training materials repositories on Glittr.org. This can easily be done using the API endpoints (https://github.com/sib-swiss/glittr#api) and applying suitable filters. For example, the data for the graphs in Fig 2 are all extracted from the Glittr.org API with the R packages httr2 [25]. The

Bioschemas endpoint is a special API endpoint in Glittr.org. It provides access to the metadata of all the repositories in Glittr.org in Bioschemas TrainingMaterial markup. This allows portals that can ingest Bioschemas markup to re-use Glittr.org data. For example, ELIXIR TeSS [8], currently ingests this metadata (https://tess.elixir-europe.org/content_providers/glittr-org).

Conclusion

When we created Glittr.org, our aim was to create a simple-to-use, web-based listing of bioinformatics training materials available on GitHub/GitLab because as training providers we saw the need for such a tool. Within one year, more than 500 repositories with bioinformatics training materials with a diverse set of topics have been included in Glittr.org. By focussing only on git repositories, Glittr.org can automatically display relevant information like popularity, license and recency that would otherwise not be possible. The resource has the potential to help in training material selection, decision-making by both students and trainers and training material database integration, as illustrated by the five use cases presented. All the information is available programmatically via the API and Bioschemas markup thereby enabling other training registries to freely update their existing entries or add entries not yet present in their database. This will further increase the coverage of these registries and promote the findability, accessibility, interoperability, and re-usability (FAIRness) of training materials in the life sciences.

References

1. Filiposka S, Green D, Mishev A, Kjorveziroski V, Corleto A, Napolitano E, et al. Draft Methodology for FAIR-by-Design Learning Materials. 2023 [cited 11 Apr 2024]. Available: https://zenodo.org/records/7875541

2. Haase R, Tischer C, Bankhead P, Miura K, Cimini B. A Call for FAIR and Open-Access Training Materials to advance Bioimage Analysis. Open Science Framework; 2024 Mar. doi:10.31219/osf.io/2zgmc

3. Alexandra Holinski, Alexia Cardona, Alice Matimba, Allegra Via, Anastasios Anastasios Papaioannou, Anna Swan, et al. ELIXIR fair training handbook. 2024 [cited 22 Mar 2024]. doi:10.5281/ZENODO.10839551

4. Garcia L, Batut B, Burke ML, Kuzak M, Psomopoulos F, Arcila R, et al. Ten simple rules for making training materials FAIR. PLoS Computational Biology. 2020;16: 1–9. doi:10.1371/journal.pcbi.1007854

5. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. doi:10.1038/sdata.2016.18

6. Sparks SD. Open Educational Resources (OER): Overview and Definition. Education Week. 12 Apr 2017. Available: https://www.edweek.org/teaching-learning/open-educational-resources-oer-overview-and-definition/2017/04. Accessed 5 Feb 2024.

7. Corpas M, Jimenez RC, Bongcam-Rudloff E, Budd A, Brazas MD, Fernandes PL, et al. The GOBLET training portal: a global repository of bioinformatics training materials, courses and trainers. Bioinformatics. 2015;31: 140–142. doi:10.1093/bioinformatics/btu601

8. Beard N, Bacall F, Nenadic A, Thurston M, Goble CA, Sansone S-A, et al. TeSS: a platform for discovering life-science training opportunities. Bioinformatics. 2020;36: 3290–3291. doi:10.1093/bioinformatics/btaa047

9. Gray AJG, Goble C, Jimenez RC. Bioschemas: From Potato Salad to Protein Annotation. ISWC 2017 Posters & Demonstrations and Industry Tracks: Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017). RWTH Aachen University; 2017. Available: https://research.manchester.ac.uk/en/publications/bioschemas-from-potato-salad-to-protein-annotation

10. Castro LJ, Palagi PM, Beard N, Attwood TK, Brazas MD. Bioschemas training profiles: A set of specifications for standardizing training information to facilitate the discovery of training programs and resources. PLOS Computational Biology. 2023;19: e1011120. doi:10.1371/journal.pcbi.1011120

11. Dogucu M, Çetinkaya-Rundel M. Tools and Recommendations for Reproducible Teaching. Journal of Statistics and Data Science Education. 2022;30: 251–260. doi:10.1080/26939169.2022.2138645

12. Wilson G. Software Carpentry: lessons learned. F1000Research; 2016. doi:10.12688/f1000research.3-62.v2

13. Hiltemann S, Rasche H, Gladman S, Hotz H-R, Larivière D, Blankenberg D, et al. Galaxy Training: A powerful framework for teaching! PLOS Computational Biology. 2023;19: e1010752. doi:10.1371/journal.pcbi.1010752

14. Ison J, Kalaš M, Jonassen I, Bolser D, Uludag M, McWilliam H, et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics. 2013;29: 1325–1332. doi:10.1093/bioinformatics/btt113

15. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26: 1112–1118. doi:10.1093/bioinformatics/btq099

16. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37: W170–W173. doi:10.1093/nar/gkp440

17. Ison J, Rapacki K, Ménager H, Kalaš M, Rydza E, Chmura P, et al. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Res. 2016;44: D38–D47. doi:10.1093/nar/gkv1116

18. Wikidata. [cited 9 Apr 2024]. Available: https://www.wikidata.org/

19. Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST Principles for digital repositories. Sci Data. 2020;7: 144. doi:10.1038/s41597-020-0486-7

20. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available: https://www.R-project.org/

21. Xie Y. bookdown: Authoring Books and Technical Documents with R Markdown. Boca Raton, Florida: Chapman and Hall/CRC; 2016. Available: https://bookdown.org/yihui/bookdown

22. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20: 631–656. doi:10.1038/s41576-019-0150-2

23. ELIXIR. Research Data Management Kit. A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075). Available: https://rdmkit.elixir-europe.org

24. Jenkins J. English as a Lingua Franca in the International University: The Politics of Academic English Language Policy. Routledge; 2013.

25. Wickham H. httr2: Perform HTTP Requests and Process the Responses. 2023. Available: https://httr2.r-lib.org