Assessing bioinformatics software annotations: bio.tools case-study

Ulysse LE CLANCHE[1], Sarah COHEN BOULAKIA[2], Yann LE CUNFF[1], Olivier DAMERON[1] and Alban GAIGNARD[3,4]

[1] Université Rennes, Inria, CNRS, IRISA—UMR 6074, Rennes 35000, France
[2] Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91405, Orsay, France
[3] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[4] IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

Corresponding author: ulysse.le-clanche@irisa.fr

Keywords Ontologies, Semantic annotations, FAIR bioinformatics software, EDAM, bio.tools

Abstract *Reproducibility and reuse of digital bioinformatics resources are essential for the development of open and cumulative science, in line with FAIR principles. To search and reuse bioinformatics tools, scientists need to be confident enough with the reliability of their annotations. Our study focuses on the quantitative and qualitative evaluation of semantic annotations in the bio.tools registry, which serves more than 30,000 bioinformatics tool descriptions, annotated with the EDAM ontology. In this work we propose to study how the EDAM ontology is used to categorize software based on scientific disciplines and the kind of data processing they allow. We also evaluate how qualitative are the annotations based on Shannon entropy. We emphasize that a particular attention should be given to the whole set of inherited annotations, from the used ontology. Our results underline the need for automatic tools to support annotation curation, reducing the annotation cost for domain experts. This study is a preliminary work aimed designing novel annotation approaches based on the combination of knowledge graphs and large language models towards more findable and reusable bioinformatics tools.*

Introduction

Ensuring reproducibility in data-driven sciences is critical for the continuous development of open and cumulative sciences. In line with the FAIR principles, this requires for digital scientific resources to be openly accessible and reusable by a wide community of researchers [1,2].

Many registries have been developed to facilitate the discovery and reuse of digital scientific resources. For instance, Zenodo and Dataverse enable the sharing of datasets and increase their discoverability through significant amount of descriptive metadata. These metadata rely on ontologies and generic controlled vocabularies such as Schema.org, DCTerms, or DCAT [3,4,5]. However, the scope of these metadata is generally limited to attribution, citation, or licensing information. They are not sufficient for searching a set of resources annotated with precise concepts, specific to certain scientific disciplines, such as "mobile genetic elements", or "protein-protein interactions".

In the field of life sciences, research communities have developed specialised registries dedicated to training materials (e.g. TeSS [6]), software tools (e.g. bio.tools [7]), or analysis pipelines (e.g. WorkflowHub [8]). These registries rely on EDAM [9], an ontology aimed at improving interoperability in bioinformatics by formally defining the nature and format of data produced and managed, different kinds of data processing, as well as the associated scientific disciplines. This ontology enables,

for example, the retrieval of algorithms dedicated to analyse a specific type of data, as well as relevant training materials. Beyond data findability, semantic indexing allows for the development of computational approaches aimed at assisting scientists in workflow composition [10] or data annotation [11]. These registries, with their growing adoption and extensive collection of resources (e.g., 30k+ bioinformatics entries in bio.tools), are key to address FAIRification challenges.

However, researchers lack insights into the reliability of their annotations. For instance, is a bioinformatics software tool sufficiently annotated? Are the chosen terms precise enough with respect to the terms hierarchy of the domain ontology? To further promote the usage of these domain-specific annotations, we need a detailed quality assessment. In this paper, we address the following question: **What is the quality of semantic annotations associated to bio.tools bioinformatics software ?**

For assessing the quality of annotations, a gold standard is required, but it does not exist yet for bioinformatics software. One approach would then consist in measuring the quality of an annotation according to its rarity: a specific annotation would be less frequent and more informative than a generic annotation, that could be assigned to a large collection of softwares.

Our main contributions are i) a characterisation of the usage of the EDAM ontology when annotating a large collection of bioinformatics software and ii) an evaluation of the specificity of the annotations through the Shannon entropy metric.

Motivating use case

Here we present a small example with two tools to illustrate the EDAM ontology's term hierarchy and its impact on tool search. We selected *Qiime2* [12] and *Vsearch* [13] as two reference bioinformatics tools used in metagenomics data analysis. *Qiime2* is annotated with topics {*Microbial ecology*, *Phylogeny*, *Metatranscriptomics*, *Metagenomics*}, and *Vsearch* with topics {*Metagenomics*, *Sequence analysis*}. The two tools share only one directly assigned annotation {*Metagenomics*}, accounting for 16% of all direct annotations, which is relatively low given their use in the same application domain.

Figure 1 shows the topic annotations for these two tools, as declared in bio.tools, along with inferred annotations from the EDAM class hierarchy. Shared topics between the tools are highlighted by a red border. There are 7 shared topics out of 17 total annotations from the combined sets, representing 41% of the annotations. This highlights the importance of considering inferred annotations when retrieving tool's annotations. This illustrates that some tools have few annotations, and have direct annotations with various levels of precision.

Material and methods

Bio.tools dataset. We leverage the bio.tools registry which now categorizes 30k+ bioinformatics tools using the EDAM ontology. In this work, we rely on the bio.tools RDF metadata available as of January 5, 2025 [6]. Based on the collected metadata, we used SPARQL queries and Python scripts to compute statistics on tool annotations [7]. From the extracted version, there are 30,025 tools described in bio.tools.

6. Available at: https://github.com/research-software-ecosystem/content/blob/master/datasets/bioschemas-dump.ttl
7. Available at: https://github.com/ulysseLeclanche/Abso_bio-tools

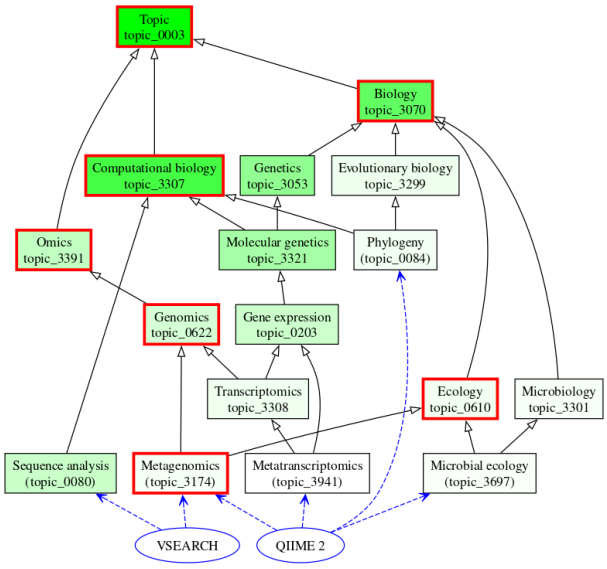

Fig. 1. Direct and inferred EDAM annotations for the scientific topic of Qiime2 and Vsearch, two metagenomics tools. Dotted blue arrows indicate direct topics, while solid black arrows represent inferred topics inherited through the EDAM hierarchy. The topics shared by both tools are highlighted by a red border. Topics colors saturation is proportional to the number of tools annotated by the topics. In this example, the dark orange boxes are shared by two tools.

EDAM. The EDAM ontology is structured into four main branches covering bioinformatics *Operation*, *Data*, *Format* and *Topic* at different levels of precision. We used EDAM version 1.25, and focused only on *Operation* and *Topic* annotations, the most used annotations in bio.tools (100k+ topics, 70k+ operations, 11k+ data and 10k+ formats). On the extracted version of bio.tools, there are 98,870 topic annotations for 29,616 tools with at least one topic, and 68,886 operation annotations for 28,299 tools with at least one operation. We identified 258 unique topics and 527 unique EDAM operations used on bio.tools.

Entropy. We used Shannon entropy as an information measure to quantify annotation quality, as it takes into account annotation rarity and distribution of EDAM annotation. A low entropy for a tool indicates either that the annotations is general, or that is annotates few tools. A high entropy indicates a balanced distribution in the attribution of annotations to the tool and more specific annotations. For a tool, topic and operation entropy are respectively the sums of their annotation entropy values.

Results and Discussion

Topics and operations in bio.tools: basic statistics and curation needs

Basic statistics were computed on the whole bio.tools dataset, comprising 30,025 software descriptions. Figure 2 shows the distribution of the number of tools annotated with a given number of EDAM *topics* or *operations*, considering both directly assigned annotations and inherited annotations. Figure 2 shows a very narrow distribution for directly assigned annotations, with a discrete distribution. This suggests that some annotations are assigned in a very standardized way and are concentrated around a small number of topics and operations. Taking inherited annotations into account yields a wider and more continuous distribution. We observe that the mean number of directly annotated

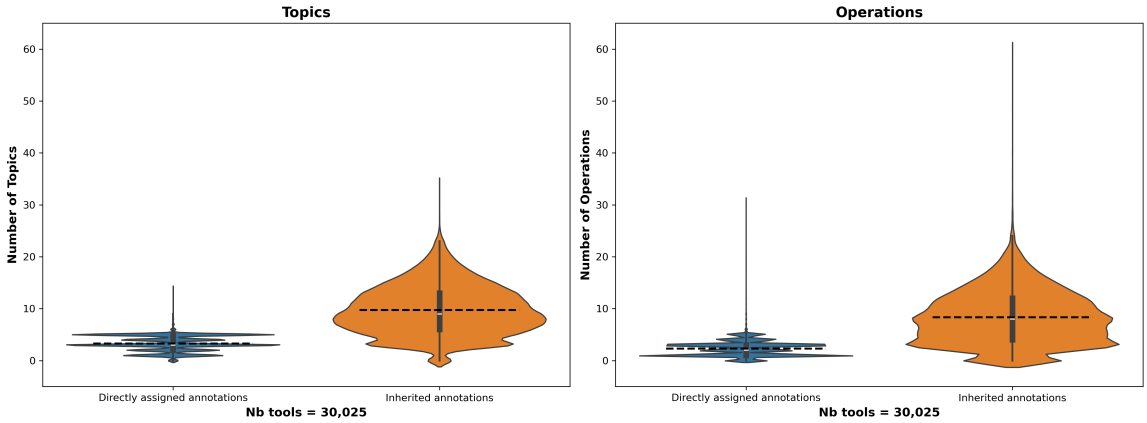

Fig. 2. Distribution of the number of tools in bio.tools according to number of topics and operation. Two conditons are tested: with direct assigned annotations, and with inherited annotations.

topics is 3.29 $\pm$ 1.45, which is comparable for operations with a means equals to 2.29 $\pm$ 1.43. When inherited annotations (ancestors) are taken into account, the mean number of topics rises to 9.73 $\pm$ 4.71 and to 8.31 $\pm$ 5.12 for operations. The EDAM hierarchy of terms helps to enrich assigned annotations for all tools, taking ancestors into account. These numbers show that the a particular attention should be given to the hierarchy of ontologies classes and not only classes typically used at resource annotation time.

Based on this dataset, we evaluated that 1,965 tools (6.54%) are not annotated with EDAM topic or operation, clearly showing the need for involving user communities to better annotate bioinformatics software. We also computed the number of tools annotated with redundant EDAM classes. For example, the magnet tool has two direct topic annotations: *Protein interactions* and *Molecular interactions, pathways and networks*. These two annotations share the same branch since *Molecular interactions, pathways, and networks* is a subclass of *Protein interactions*. Adding the *Protein interactions* annotation does not provide any additional information, as it is inherited from the ontology. We estimated that 3,405 tools (11.34%) have redundant direct topic annotations, and 2,055 tools (6.84%) have redundant direct operation annotations. This highlights the need for better curation in the tools database. Finally, there are 1,114 deprecated annotations, 54 tools with at least 1 deprecated topic and 347 tools with at least 1 deprecated operation, also highlighting the need for database curation.

Are bioinformatics software annotated with informative enough classes ?

We computed the entropy of tools annotated with EDAM *topics* and *operations*. By only considering direct annotations, the mean *topic* entropy is 0.57 $\pm$ 0.32, but when considering inherited classes, the entropy grows to 2.98 $\pm$ 1.72. We observed the same increase for *operations* with 0.23 $\pm$ 0.18 for direct annotations and 2.01 $\pm$ 1.32 for inherited ones. The entropy of topics for direct and inherited annotations is greater than the entropy of operations. This reflects both a greater diversity in the assignment of topic annotations compared to operation annotations and the size difference of these two branches. We calculated a Pearson correlation coefficient of 0.96 showing a positive correlation between the number of annotations and their entropy. The current entropy measurement can only be increased by adding annotations, even if the annotations are not very informative. However, if we

take two tools with a similar number of annotations but different rarity in term annotation, the tool with the most rare terms will have a higher entropy.

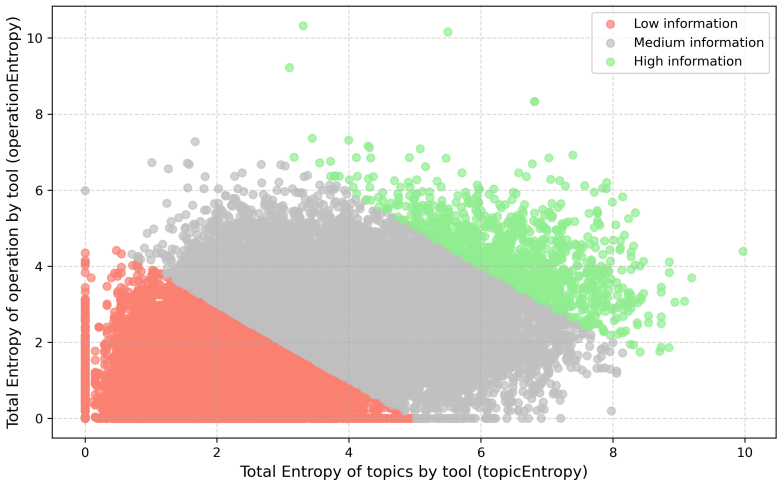

Fig. 3. Distribution of topic and operation entropy for bio.tools software taking into account inherited classes. The total entropy E_{to} of a tool is the sum of the topic entropy and the operation entropy. Red dots represent tools with low information ($SumEnt < 5$), gray dots represent tools with medium information $5 <= SumEnt < 10$, and green dots show tools annotated with highly informative annotations ($SumEnt >= 10$).

Figure 3 shows how informatively tools are annotated with inherited classes, considering both the *topic* entropy and the *operation* entropy. Tools are grouped into three categories based on an arbitrary threshold on the sum of these two metrics (max $SumEnt = 15.65$). The majority of tools - 29,061 (96.78%) - belong to the low or medium information categories. A few tools (964) have an entropy sum greater than 10, indicating a high information level for their annotations. 53.05% of tools (in red), i.e. 15,929 tools, are annotated with a low level of information, suggesting that they should be prioritized for database curation.

Increasing the number of annotations has a positive impact on the quality of tool information. The distribution of tools with redundant annotations is similar in each group, with 2,369 (14.87%) redundant tools in the low information level, 2,193 (16.70%) in the medium level and 86 (8.92%) in the high level. However, among the top 10 tools with the highest entropy sum, 60% of tools have redundant annotations. Redundant annotations should be removed, as they artificially increase entropy. A metric that penalizes, more than entropy, annotation generality and redundancy would be interesting.

Conclusion

In this work, we have shown that considering inherited annotations from the ontology increases the number of annotations for topics and operations in the whole set of tools, making the tools more searchable and reusable. To quantify annotation quality, we used Shannon entropy, which takes into account annotation rarity and the distribution of EDAM annotations. This measure enabled us to compare the annotation quality of the tools with each other, and to identify a set of 15,929 tools (53.05%) with a low level of information annotations. This set of tools should be prioritized for future database curation activities. The main limit of our approach is the lack of ground-truth to assess the accuracy of annotations, Shanon entropy assess the rarity of annotations, which does not mean that

they are correct. To address this issue, we are currently working with bioinformatics experts to define a reference dataset of highly curated annotations. Through this study, we have also seen the impact of EDAM ontology evolution on annotation quality, with the identification of redundant or obsolete annotations.

This opens for new research directions we will pursue as future works. We are currently working on implementing more suited metrics to better assess the quality of EDAM annotations. To support curation tasks, we aim at combining large language models and knowledge graphs [14] as a means to suggest more informative annotations, or to identify possibly missing classes in the ontology. Although this work is grounded to bio.tools, it aims at being generalized to other application domains also using ontologies and registries for annotating and sharing FAIR digital resources.

Availability and Implementation

All the code for extracting metadata from the RDF schema, creating article figures and calculating tool annotation statistics is available on the following github repository: https://github.com/ulysseLeclanche/Abso_bio-tools.

Funding information

This work is supported by the Agence Nationale de la Recherche under the France 2030 program, ANR-22-PESN-0007 ShareFAIR.

References

[1] Kinkade D, Shepherd A. Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. Geoscience Data Journal. 2022;9(1):177-86.

[2] Top J, Janssen S, Boogaard H, Knapen R, Şimşek-Şenel G. Cultivating FAIR principles for agri-food data. Computers and Electronics in Agriculture. 2022;196:106909.

[3] Guha RV, Brickley D, Macbeth S. Schema.org: Evolution of Structured Data on the Web. Queue. 2015;13:10 37. Available from: https://api.semanticscholar.org/CorpusID:27038003.

[4] Weibel SL, Kunze JA, Lagoze C, Wolf M. Dublin Core Metadata for Resource Discovery. RFC. 1998;2413:1-8. Available from: https://api.semanticscholar.org/CorpusID:43249830.

[5] Archer P. Maali F, Erickson J, editors. Data Catalog Vocabulary (DCAT) (W3C Recommendation); 2014. Online. Available from: https://www.w3.org/TR/vocab-dcat/.

[6] Beard N, Bacall F, Nenadic A, Thurston M, Goble CA, Sansone SA, et al. TeSS: a platform for discovering life-science training opportunities. Bioinformatics. 2020 02;36(10):3290-1. Available from: https://doi.org/10.1093/bioinformatics/btaa047.

[7] Ison JC, Ienasescu H, Chmura P, Rydza E, Ménager H, Kala M, et al. The bio.tools registry of software tools and data resources for the life sciences. Genome Biology. 2019;20. Available from: https://api.semanticscholar.org/CorpusID:199538589.

[8] Gustafsson J, Wilkinson SR, Bacall F, Pireddu L, Soiland-Reyes S, Leo S, et al. WorkflowHub: a registry for computational workflows. ArXiv. 2024;abs/2410.06941. Available from: https://api.semanticscholar.org/CorpusID:273228446.

[9] Ison JC, Kala M, Jonassen I, Bolser DM, 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. Available from: https://api.semanticscholar.org/CorpusID:1626822.

[10] Kasalica V, Schwämmle V, Palmblad M, Ison J, Lamprecht A. APE in the Wild: Automated Exploration of Proteomics Workflows in the bio.tools Registry. Journal of proteome research.

2021;20(4):2157–2165. Publisher Copyright: © 2020 American Chemical Society. All rights reserved.

[11] Gaignard A, Skaf-Molli H, Belhajjame K. Findable and reusable workflow data products: A genomic workflow case study. Semantic Web – Interoperability, Usability, Applicability. 2020 May:1-13. Available from: https://hal.science/hal-02903805.

[12] Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science; 2018. .

[13] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4.

[14] Gilbert S, Kather JN, Hogan A. Augmented non-hallucinating large language models as medical information curators. NPJ digital medicine. 2024;7(1):100.