FAIR-Checker : Checking and Inspecting metadata for FAIR bioinformatics resources

Thomas Rosnet, F de Lamotte, Marie-Dominique Devignes, Vincent Lefort, A Gaignard

FAIR-Checker : Checking and Inspecting metadata for FAIR bioinformatics resources

T. Rosnet[1,5], F. de Lamotte[1,6], M.-D. Devignes[1,3], V. Lefort[1,2], A. Gaignard[1,4].

[1]Institut Français de Bioinformatique, CNRS UAR 3601, France, [2]LIRMM, Univ Montpellier, CNRS, Montpellier, France, [3]LORIA, Université de Lorraine, CNRS, Inria, Nancy, France, [4]L'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France, [5]TAGC/INSERM U1090, Univ Aix-Marseille, Marseille, France, [6]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France

• *Open Sciences* • *FAIRMetrics* • *Knowledge Graphs* • *Linked Data* • *FAIR Data* • *RDF / SPARQL / SHACL*

1. Introduction

The continuous increase of life science data production raises the importance of better sharing and reusing biological digital resources (datasets, bioinformatics tools or workflows, training materials, etc.). This led to an increasing importance of Open Science and Reproducibility, **requiring rich and machine-actionable metadata**. To that end, FAIR principles [1] have been proposed and are being adopted by large communities. However, **assessing how much a resource is FAIR is nowadays challenging** since answering human-oriented questionnaires is time-consuming and computational evaluations (FAIRMetrics, RDA Maturity Indicators) often require technical expertise.
In this work, we propose a new release of **FAIR-Checker**, aimed at empowering scientists and developers in FAIRifing their resources.

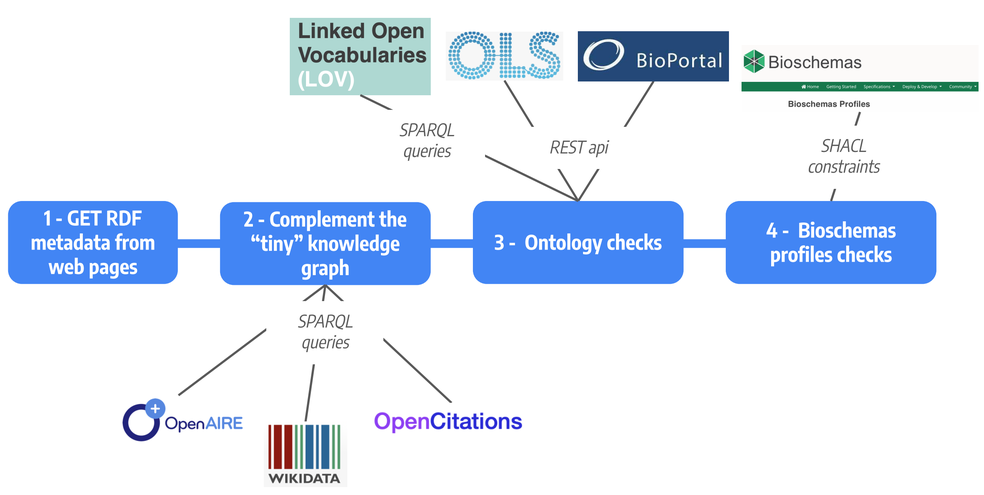
2. Motivating use cases

1. I am a **data producer**, I published my dataset through an online registry, does it provide rich metadata ?
   - Are these metadata **interoperable, reusable** ?
   - Is the registry exposing metadata through a community agreed controlled vocabulary ?
2. I am a **software developer**, my source code is on GitHub, but not mature enough to be part of a registry yet.
   - Is my resource exposing agreed and widly used ?
   - Am I missing **required** or **recommended** metadata ?
3. I am in **charge of a bioinformatics registry** such as Bio.tools [2] which provides online access to many tools descriptions.
   - How FAIR are the metadata exposed in each tool page ?
   - Inspect if metadata are compliant with the specific profile for their community.
   - What can be done to improve the quality of exposed metadata ?

3. Approach

A. Check. We propose a web interface[a] aimed at empowering scientists to progress in the FAIRification of their resources through a global assessment and technical recommendations. This tool supports an iterative process, leveraging the **web semantic technologies (RDF, SPARQL)** and metrics-specific guidelines, with references to the FAIR Cookbook and RDMkit initiative.

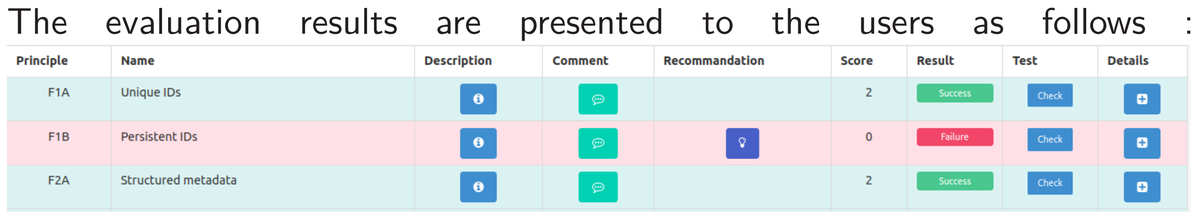
B. Inspect. We use semantic technologies to help users in providing fine-grained community-agreed metadata. We assemble a Knowledge Graph from embedded RDF, complemented by public SPARQL endpoints. We check that used ontology terms are already known in reference registries (LOV, OLS, BioPortal). Bioschemas specifications are used to generate SHACL shapes. Their evaluation informs users on missing metadata, required or recommended for specific resources (genes, proteins, training, tools, etc.).



a. https://fair-checker.france-bioinformatique.fr

4. Results

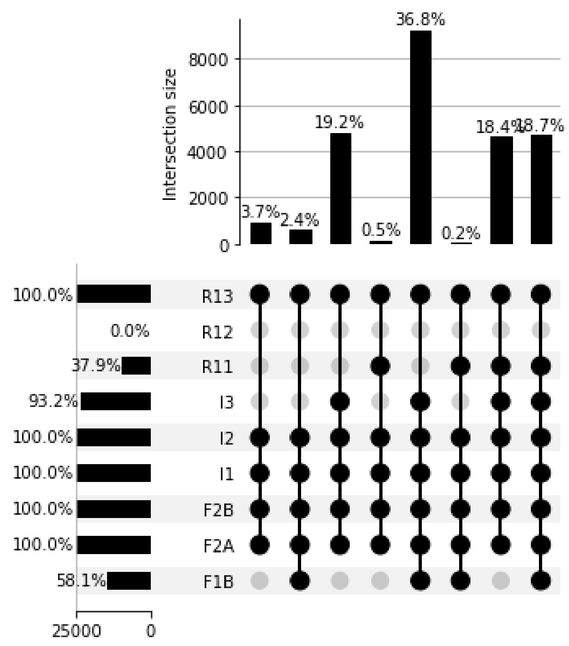
A. Check.
The evaluation results are presented to the users as follows :



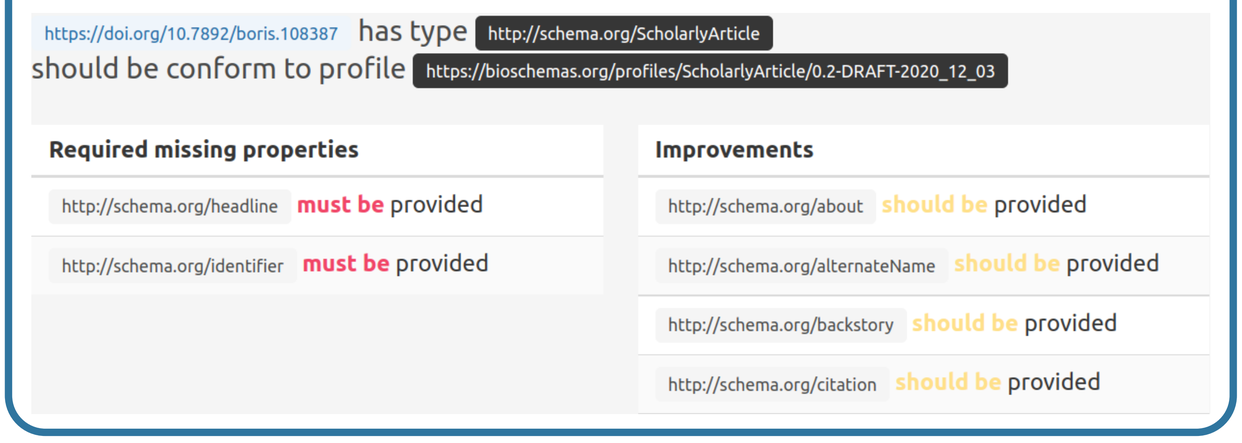
Principle	Name	Description	Comment	Recommandation	Score	Result	Test	Details
F1A	Unique IDs				2	Success	Check	
F1B	Persistent IDs				0	Failure	Check	
F2A	Structured metadata				2	Success	Check	

How FAIR are Bio.Tools registered softwares ?

After validating more than 25,000 tool pages from Bio.tools we plotted the proportion of tools complying with the FAIR principles. This allows to indicate where the metadata quality can be improved by better annotations. Only 37,9% of the tools have a Licence. More than 60% of the tool descriptions should be improved by providing a License.



B. Inspect.
We show to the user which properties in the metadata can be added to improve the FAIRness of the resource based on the corresponding Bioschemas profile.



5. Future works

A publication of this work is currently under review. As future work, we aim (i) to support HTTP content-negotiation for web sites not embedding metadata in their web pages, (ii) to enhance the matching between a resource type and the corresponding Bioschemas profile, and (iii) to allow the user to annotate and complete its missing metadata. In addition, we plan to develop an API for a better scalability and interoperability of FAIR-Checker.

6. References

[1] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[2] J. Ison, H. Ienasescu, P. Chmura, E. Rydza, H. Ménager, M. Kala, V. Schwämmle, B. Grüning, N. Beard, R. Lopez, S. Duvaud, H. Stockinger, B. Persson, R. S. Vaeková, T. Raek, J. Vondráek, H. Peterson, A. Salumets, I. Jonassen, R. Hooft, T. Nyrönen, A. Valencia, S. Capella, J. Gelpí, F. Zambelli, B. Savakis, B. Leskoek, K. Rapacki, C. Blanchet, R. Jimenez, A. Oliveira, G. Vriend, O. Collin, J. van Helden, P. Løngreen, and S. Brunak. The bio.tools registry of software tools and data resources for the life sciences. *Genome Biol*, 20(1) :164, 08 2019.