

FAIR-Checker – Supporting the findability and reusability of digital life science resources

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

¹TAGC/INSERM U1090, Univ Aix-Marseille, Marseille, France, ²LIRMM, Univ Montpellier, CNRS, Montpellier, France, ³LORIA, Université de Lorraine, CNRS, Inria, Nancy, France, ⁴L'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France, ⁵Institut Français de Bioinformatique, CNRS UMS 3601, France, ⁶UMR 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

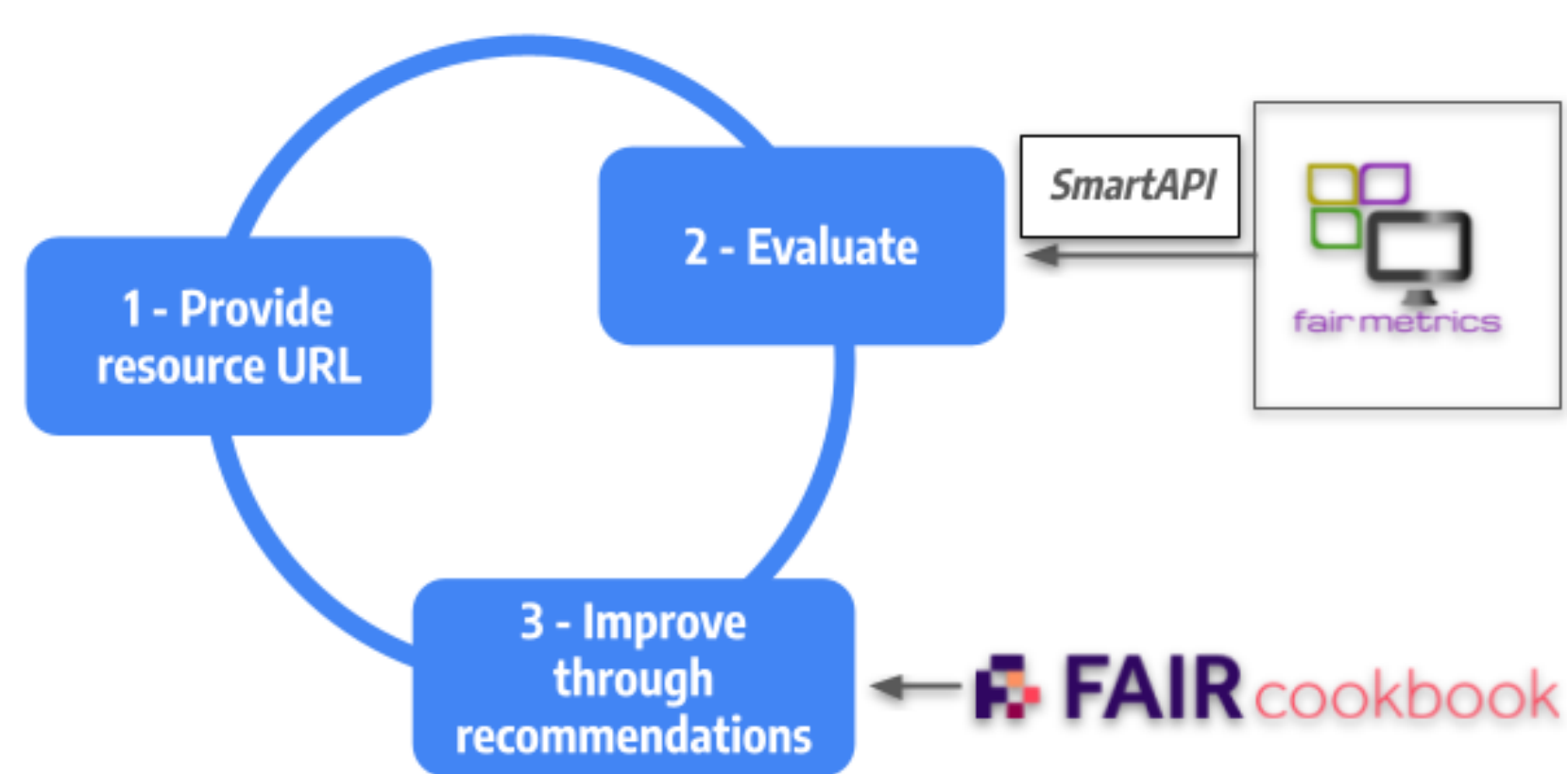
Digital disruption and data deluge in sciences. The explosion of web-accessible digital resources makes the discovery of meaningful datasets, softwares, methods, training, challenging for scientists. **Providing machine-readable metadata is the cornerstone to scale up the development and adoption of Open Sciences.** It is urgent to better share and reuse biological digital resources. FAIR principle are currently being adopted by many scientific communities. However, **assessing how much a resource is FAIR is nowadays challenging.** Answering human-oriented questionnaires is time-consuming and computational evaluations (FAIRMetrics [1], RDA Maturity Indicators) often require technical expertise. In this work, we aim at empowering scientists and developers in FAIRifying their resources from the very early stages.

2. Motivating use cases

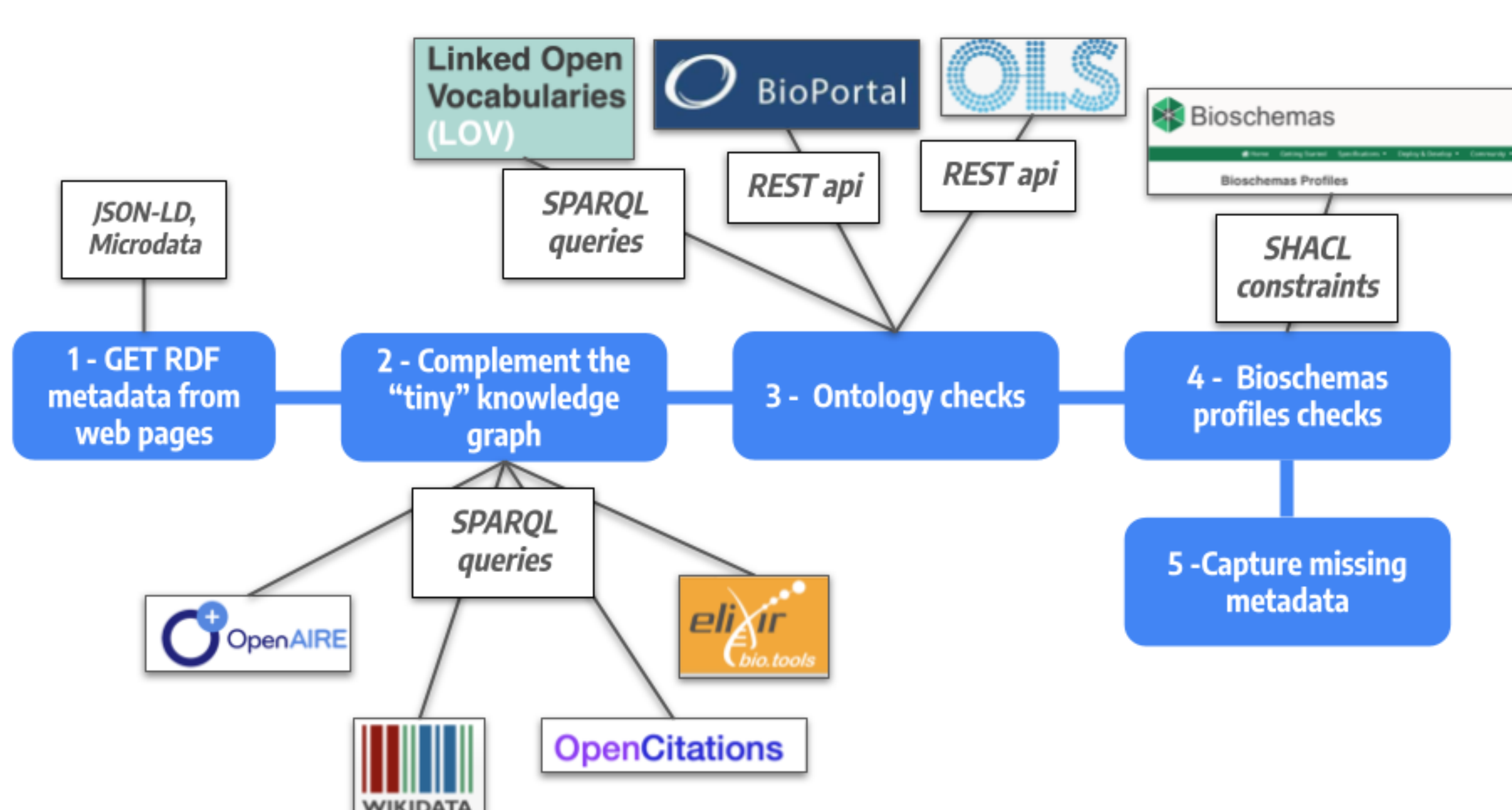
- I am a **data producer**, I published my dataset through an online registry, does it provide rich metadata?
 - Are these metadata **interoperable, reusable**?
 - Is the registry exposing metadata through a community agreed controlled vocabulary?
- I am a **software developer**, my source code is on GitHub, but not mature enough to be part of a registry yet.
 - How to check if my tool is **findable** enough?
 - Which kind of metadata should I advertise?
 - Am I missing **required** or **recommended** metadata?
- I am in **charge of a bioinformatic platform** which provides online available services such as PhyML (<http://www.atgc-montpellier.fr/phyml/>). PhyML[2] is already registered in bio.tools.
 - What can I do to improve the search engine indexing of PhyML?
 - How to improve PhyML **findability**?
 - Which metadata should I use to improve PhyML description?

3. Approach

A. FAIRMetrics. We propose a web interface (<https://fair-checker.france-bioinformatique.fr>) 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 FAIRMetrics APIs and metrics-specific guidelines provided by the FAIR Cookbook initiative.



B. Metadata quality checks. 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. 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.). Finally, it results in a form to gather and enhance metadata.



4. Results

A.1 Evaluation overview

Improving PhyML FAIRness with metadata provided by bio.tools.



A.2 Metric-specific recommendations

Principle	Name	Description	Time	Comment	Recommendation	Score	Result	Test
F1	Unique Identifier		0:00:02			1	Success	Check
F2	Structured Metadata		0:00:15			1	Success	Check
F2	Grounded Metadata		0:00:17			1	Success	Check
F3	Data Identifier Explicitly In Metadata		0:00:18		Ensure that meta-data describing your resource use the machine readable standards parsed by major search engines such as schema.org OpenGraph, etc. Also, see the corresponding FAIR Cookbook page: https://fairplus.github.io/the-fair-cookbook/content/recipes/findability/seo.html	0	Failure	Check
F3	Metadata Identifier Explicitly In Metadata		0:00:15			0	Failure	Check
F4	Searchable in major search engine		0:00:18			0	Failure	Check

B. Metadata quality improvement

Properties

http://edamontology.org/has_input	
http://edamontology.org/has_output	
http://schema.org/author	
http://schema.org/citation	
http://schema.org/description	

Improvements

- Property <http://schema.org/applicationCategory> **should be** provided
- Property <http://schema.org/softwareVersion> **should be** provided

5. Future works

We aim at extending our tool to (i) support multiple resource types in line with the different released Bioschemas profiles and (ii) provide a common and synthetic view on other FAIR recommendations such as the RDA maturity indicators, as well as the forthcoming EOSC FAIR metrics.

6. References

- 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.
- Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3) :307–321, 05 2010.