FAIR-Checker: graphes de connaissances et technologies sémantiques pour la découvrabilité et la réutilisation des ressources scientifiques en ligne

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

[1] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[2] TAGC/INSERM U1090, Univ Aix-Marseille, Marseille, France
[3] Institut Français de Bioinformatique, CNRS UAR 3601, France
[4] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France
[5] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[6] Université de Lorraine, CNRS, Inria, LORIA, F-54500 Nancy, France

alban.gaignard@univ-nantes.fr

Résumé

Les principes FAIR ont été adoptés par de larges communautés scientifiques pour répondre aux enjeux des sciences plus reproductibles et plus ouvertes. Cependant, leur évaluation reste difficile, car elle nécessite du temps et de l'expertise technique. Pour répondre à ces enjeux d'évaluation, nous proposons FAIR-Checker. Cet outil web propose deux facettes : un module "Check" qui permet d'évaluer les principes FAIR sur les métadonnées de ressources web, nécessitant peu de connaissances préalables, et un module "Inspect" qui aide les développeurs à améliorer la qualité des métadonnées et donc la FAIRification de leur ressource. FAIR-Checker s'appuie sur les technologies du web sémantique telles que les requêtes SPARQL et les contraintes SHACL. Nous proposons une évaluation expérimentale de FAIR-Checker via l'analyse de plus de 25 000 descriptions de logiciels bioinformatiques.

Mots-clés

FAIR, Schema.org, Bioschemas, SPARQL, SHACL, Open Science

1 Introduction

Dans un contexte de production massive de données scientifiques, les sciences ouvertes sont actuellement perçues par de nombreuses communautés scientifiques et organismes de financement de la recherche comme un moyen de faciliter la réutilisation de données, le développement de méta-analyses, ou bien encore d'améliorer la puissance statistique et la validité des modèles scientifiques. Les principes FAIR [1] ont été introduits en 2016 pour faciliter l'accès, la découvrabilité, l'interopérabilité et la réutilisation de ressources scientifiques numériques. Au départ, l'évaluation de ces critères se faisait principalement par questionnaires. Assez rapidement, des outils automatiques [2, 3, 4] ont été développés [1] [2] [3].

Malgré ces outils, il est toujours difficile aujourd'hui, pour les producteurs de données ou les développeurs d'entrepôts, de choisir et de mobiliser les ressources sémantiques déjà existantes dans leur démarche de FAIRification.

Dans cet article, nous proposons FAIR-Checker, un outil, s'appuyant sur les graphes de connaissances et les technologies du web sémantique, dans le but de rendre les producteurs et les développeurs de ressources scientifiques en ligne plus autonomes dans la mise en oeuvre des principes FAIR. Les principales contributions sont *i)* une collection de requêtes SPARQL visant à évaluer les principes FAIR, *ii)* un générateur de contraintes SHACL visant à évaluer les profils de métadonnées et ainsi améliorer l'exhaustivité de métadonnées, et *iii)* une évaluation de notre approche sur des ressources logicielles en bioinformatique.

2 Approche

L'idée générale de FAIR-Checker est de promouvoir l'utilisation de métadonnées intégrées dans les pages Web afin de faciliter la recherche et la réutilisation des ressources scientifiques en ligne. La figure 1 décrit le processus de collecte, d'enrichissement et d'analyse des annotations sémantiques tout en mobilisant les graphes de connaissances publics.

A partir d'une URL de page web, la première étape consiste à extraire des annotations sémantiques (JSON-LD, RDFa ou microdata) (❶). Cela constitue un graphe de connaissances minimal (KG) qui est interrogé avec SPARQL pour l'évaluation des métriques FAIR (❷). Ensuite, pour chaque entité (*e.g. person, dataset, software, etc.*), les KG publics sont interrogés pour obtenir les triplets RDF associés (❸). L'idée est, i) pour les producteurs de données,

1. https ://fairsharing.github.io/FAIR-Evaluator-FrontEnd
2. https ://fairshake.cloud/
3. https ://www.fairsfair.eu/f-uji-automated-fair-data-assessment-tool

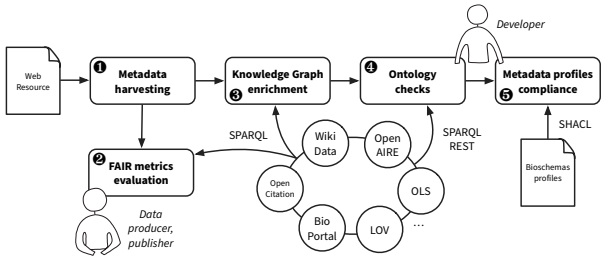

FIGURE 1 – Collecte, enrichissement et analyse des annotations sémantiques des pages web, en lien avec les principes FAIR

de limiter leurs efforts d'annotation, et ii) pour les développeurs/mainteneurs de graphes de connaissance, de les faire contribuer à la mise en oeuvre des principes FAIR. Ensuite (❹), nous évaluons si les classes ou propriétés utilisées dans les métadonnées font partie des ontologies de référence. Nous nous appuyons pour cela sur BioPortal [4] ou OLS [5], ou des registres généraux tels que LOV [6]. Enfin, comme de plus en plus de pages Web sont annotées avec Schema.org [5], nous utilisons des profils communautaires (Bioschemas [6]) pour évaluer l'exhaustivité des annotations sémantiques (❺). Ces profils Bioschemas sont automatiquement transformés en contraintes SHACL. Ces contraintes mettent en évidence les propriétés manquantes, identifiées comme obligatoires ou recommandées par la communauté pour décrire un certain type de ressource. Cela permet d'améliorer de la qualité des métadonnées.

3 Résultats

Implémentation FAIR-Checker est disponible en ligne [7] sous la forme d'une application web et d'une API REST. Le code source est disponible via GitHub [8] sous license MIT. Depuis Janvier 2023, l'outil a été utilisé intensivement et a permis l'évaluation de plus de 350 000 métriques.

Evaluation Bio.Tools [9] est une base de données répertoriant les logiciels de bioinformatique. En avril 2022, ce registre disposait de 25048 de descriptions d'outils. Ce registre a été instrumenté pour exposer des annotations sémantiques Schema.org afin d'améliorer la découvrabilité des outils. Nous avons utilisé FAIR-Checker, pour évaluer la collection complète de 25048 descriptions de logiciels bioinformatiques.

Nous avons pu observer que les entrées avec la plus grande couverture de métriques FAIR (9/11) représente 18,7% de l'ensemble de la collection. Nous avons également pu noter que 36,8% des descriptions de logiciels ont une bonne couverture des métriques FAIR (8/11) mais ne valident pas R1.1 (license). Finalement, aucune des entrées de bio.tools ne valide la métrique R1.2 (provenance). L'instrumentation de bio.tools pour permettre l'exposition de métadonnées de provenance aurait un impact direct sur plus de 25 000 outils bioinformatiques.

Sur cette même collection, nous avons évalué la conformité des descriptions de logiciels avec le profil Bioschemas *ComputationalTool*. Nous avons pu observer que toutes les propriétés identifiées dans le profil comme "obligatoires" sont bien disponibles. En revanche, bio.tools n'expose que 54% des propriétés identifiées comme "recommandées". Ces résultats méritent d'être étudiés par les développeurs du registre bio.tools afin d'améliorer encore la qualité des métadonnées des ressources logicielles.

4 Conclusion et perspectives

Dans cet article, nous présentons FAIR-Checker, un outil web qui rend les technologies sémantiques et les graphes de connaissances plus accessibles aux utilisateurs non experts, afin de développer l'usage et d'améliorer la qualité des métadonnées, contribuant ainsi à l'adoption des principes FAIR en pratique. Les premiers retours des utilisateurs ont mis en évidence des attentes importantes pour soutenir le développement des démarches "sciences ouvertes". Dans les travaux futurs, nous visons à mieux prendre en charge les techniques de négociation de contenu, plus efficacement exploiter les graphes de connaissances publics, et également calculer des distances sémantiques avec les profils de la communauté Bioschemas afin de suggérer les plus pertinents, et ainsi favoriser leur adoption.

Références

[1] Wilkinson et al. *The FAIR Guiding Principles for scientific data management and stewardship*, Sci Data 3, 2016.

[2] Wilkinson et al. *Evaluating FAIR maturity through a scalable, automated, community-governed framework*. Sci Data 6(1), 2019.

[3] Clarke et al. *FAIRshake : Toolkit to Evaluate the FAIRness of Research Digital Resources*. Cell Syst 9(5), 2019.

[4] Devaraju et al. *An automated solution for measuring the progress toward fair research data*. Patterns 2(11), 2021.

[5] Benjelloun, O., Chen, S., Noy, N. (2020). *Google Dataset Search by the Numbers*. International Workshop on the Semantic Web.

[6] Castro et al. *Bioschemas : schema.org for the life sciences*, SWAT4LS, 2017

4. Bioportal : https ://bioportal.bioontology.org
5. Ontology Lookup Service : https ://www.ebi.ac. uk/ols/index
6. Linked Open Vocabularies : https ://lov.linkeddata.es/dataset/lov/
7. https ://fair-checker.france-bioinformatique.fr
8. https ://github.com/IFB-ElixirFr/FAIR-checker
9. http ://bio.tools