Système d'évaluation des métadonnées pour la production de Data Paper : l'importance d'un contenu actionnable par la machine

T. Genthon[1], Y. Le Bras[2]

[1] Université Grenoble-Alpes, UAR PatriNat
[2] Muséum national d'histoire naturelle, UAR PatriNat
...

tanguy.genthon1@etu.univ-grenoble-alpes.fr
yvan.le-bras@mnhn.fr

Résumé

Afin de guider le chercheur à un meilleur respect des principes FAIR, l'infrastructure de recherche "Pôle national de données de biodiversité (PNDB) développe actuellement un outil de saisie de métadonnées de biodiversité et aidant à la production de Data Paper. Cet outil permettra au chercheur de pouvoir évaluer la qualité de ses métadonnées et créera une version préliminaire de Data Paper permettant de valoriser ses données et l'incitant ainsi à produire des données de qualité. Ce Data Paper pourra être édité par l'utilisateur afin de pouvoir le finaliser et de potentiellement le publier

Mots-clés

FAIR, Métadonnée, Réutilisable, Data Paper, EML, Machine Actionable.

Abstract

In order to guide the researcher to a better respect of the FAIR principles, the biodiversity PNDB research infrastructure is currently developping a tool for biodiversity metadata entry and helping to produce Data Paper. This tool will allow researchers to assess the quality of their data and create a preliminary version of a Data Paper that will allow them to add value to their metadata and thus encouraging the production of a FAIR data. This Data Paper can be converted into a file that can be edited by the user in order to finalise it and potentially publish it.

Keywords

FAIR, Metadata, Reusable, Data Paper, EML, Machine Actionable.

1 Introduction

Dans un objectif d'amélioration de la connaissance scientifique il est important de pouvoir accéder aux connaissances déjà acquises par la communauté. Ainsi, permettre la réutilisabilité de la donnée lorsqu'on la produit est primordial. Pour y parvenir, il faut que la description de ces données soit complète, précise et facilement compréhensible afin qu'une machine et/ou une personne tierce à la production de la donnée soit capable de la comprendre et de la réutiliser. De cette problématique les principes FAIR ont émergés en 2014 et ont été généralisés en 2016 [4]. Les principes FAIR s'articulent autour de 4 grands principes. Il faut que les données soient Faciles à trouver, Accessibles, Interopérables et Réutilisables. Ces principes proposent alors différents critères permettant de guider l'utilisateur lors du partage de la donnée pour la rendre plus FAIR.

La métadonnée permet d'aider à l'application de ces principes car elle est l'information donnant le contexte nécessaire à la compréhension de la donnée brute par un utilisateur externe à la création de cette donnée. Elle est également le premier vecteur d'informations sémantiques.

En effet, la métadonnée est par elle-même une source d'information sémantique notamment lorsque l'on utilise un standard spécifique à son domaine d'application comme on peut le voir avec l'utilisation de l'Ecological Metadata Language (EML) pour les données de biodiversité [2]. L'utilisation d'un tel standard permet une meilleure organisation de la connaissance pour un usage par l'humain ou la machine. L'EML est un standard de métadonnées spécialisé dans la description des données de biodiversité et permet l'intégration d'un nombre bien plus important d'information que d'autres standards comme ISO-19115 ou bien DCAT. On observe notamment un meilleur classement des ressources sémantiques telles que les thésaurus ou les annotations sémantiques. Un enjeu particulier réside ainsi dans le fait que la métadonnée soit la plus complète possible notamment pour limiter une perte d'information dans le temps et maximiser le potentiel de réutilisation. Ainsi, utiliser un format adapté au domaine et augmenter et/ou améliorer l'information au sein de ces métadonnées permet une meilleure accessibilité de l'information sémantique par la machine ou même par l'Homme. Il est donc nécessaire de s'assurer de la bonne rédaction de la métadonnée afin de pouvoir la rendre FAIR et "machine actionnable".

2 Materiels et Méthodes

L'infrastructure de recherche "Pôle national de données de biodiversité (PNDB) a notamment pour objectif de faciliter la création, la manipulation ainsi que la curation de méta-

données en permettant de fournir un compte-rendu présentant une évaluation de la qualité de la métadonnée en appliquant des critères spécifiques au standard EML. Ce score pourra alors être utilisé afin que l'utilisateur puisse corriger ses données et en améliorer au mieux le score.
L'outil s'inspire directement du projet MetaDIG [1] du NCEAS, utilisé dans le réseau mondial des données d'observation de la terre DataOne [3] pour fournir un score FAIR en fonction d'une suite de critères ("NCEAS/metadig-checks" sur GitHub) déjà présent dans DataOne. L'objectif est alors de reprendre cette méthodologie de production d'un rapport d'évaluation de la qualité des métadonnées afin de pouvoir les utiliser à travers un service d'amélioration. Cela permettrait alors de pouvoir obtenir un résumé de la qualité de la donnée avant de la rentrer dans le système afin que l'utilisateur puisse prendre compte ce compte-rendu et appliquer aux mieux les critères manquants afin que les données entrées dans le système soient plus FAIR.
Les différents critères utilisés par MetaDIG ont alors été repris, traduits sous forme de nouveau package R pour en faciliter l'utilisation postérieure et pour servir de guide afin de permettre l'amélioration du contenu des fiches de métadonnées produites par les communautés de recherche.
En parallèle, le PNDB teste le développement d'un service de création d'une version préliminaire de Data Paper via le projet OpenMetaPaper financé par le Fond National Science Ouverte. Le Data Paper est un type d'article scientifique particulier évalué par les pairs et donc citable qui peut également être vu comme une représentation de la métadonnée sous une forme éditoriale particulière. Il est ici proposé d'ajouter dans ce service de production de version préliminaire de Data Paper, le rapport d'évaluation de qualité obtenu avec les tests de la suite MetaDIG. Ce service se base sur le package R 'emldown' de rOpenSci, pour le moment archivé, en y ajoutant certains éléments supplémentaires.
Ce Data Paper pourra être converti dans un format éditable comme docx ou Markdown ou sera directemeent éditable sur l'HTML afin d'achever la création de l'article.
Enfin, cet outil a été intégré à Galaxy, une plateforme en ligne facilitant l'accès aux outils de traitement de données.

3 Résultats et Discussion

Une application R Shiny a ainsi pu être construite. Elle inclut pour le moment un affichage des métadonnées sous forme de Data Paper ainsi qu'un compte-rendu de qualité reprenant 45 tests sur les 51 utilisés pour MetaDIG. Les différents tests renvoient donc chacun un statut : "Success" si le test est validé et "Warning" ou "Failure" en fonction de l'importance de la réussite du test ainsi qu'un message expliquant le statut pour permettre à l'utilisateur de pouvoir se corriger si besoin.
Le compte-rendu contient un tableau reprenant tous les statuts et messages de l'ensemble des tests ainsi qu'une représentation d'un score de qualité global. Le score de qualité global est pour le moment calculé en fonction du nombre

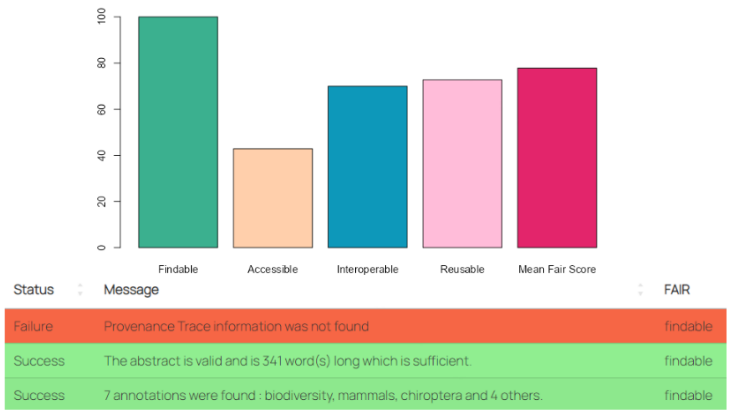

FIGURE 1 – Exemple de compte-rendu de qualité

de "Success" de l'ensemble des tests. Cependant, une pondération du score en fonction de l'importance du test est envisagée.
Le score de qualité sera également calculé pour chacune des catégories FAIR afin que l'utilisateur puisse adapter ces données en fonction de comment il souhaite la partager.
L'application produit également un visuel de Data Paper avec notamment le titre, les auteurs, les mots clés, les annotations sémantiques, le résumé, les couvertures géographiques, taxonomiques et temporelles de l'étude ainsi qu'une présentation des attributs des différents fichiers de données.
Cette fonctionnalité permet d'aider le chercheur à la production d'un Data Paper qui pourra ainsi être publié valorisant le travail de l'ensemble des contributeurs à la production de la donnée.
Un tel système représente un moyen d'encourager l'entrée d'une métadonnée détaillée de qualité dans le système. En plus de son aide à la production de Data Paper, l'outil permet de rendre beaucoup plus lisible et compréhensible une fiche de métadonnées afin de faciliter son partage.

Références

[1] Habermann, T. , MetaDIG recommendations for FAIR DataCite metadata, 2019.

[2] Matthew B. Jones, Margaret O'Brien, Bryce Mecum, Carl Boettiger, Mark Schildhauer, Mitchell Maier, Timothy Whiteaker, Stevan Earl, Steven Chong., Ecological Metadata Language version 2.2.0, *KNB Data Repository*, 2019.

[3] Matthew Jones, Peter Slaughter, and Ted Habermann, Quantifying FAIR : metadata improvement and guidance in the DataONE repository network, *KNB Data Repository*, 2019.

[4] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., « The FAIR Guiding Principles for Scientific Data Management and Stewardship », . *Scientific Data*, 3 (1) : 160018, 2016.