Improving SoBigData platform's FAIR principles through decision tree and web-based services

Payel Patra
payel.patra@univaq.graduate.it
University of L'Aquila
L'Aquila, Italy

ABSTRACT

It is essential for all infrastructures to manage scientific data with machine-actionability, a term referring to the ability of computational systems to locate, access, interact with, and reuse data with barely any human supervision. Working with open science data requires attention to both the raw data and adequate metadata, both of which must be completely reproducible. Our goal is to close this gap by putting forth a decision tree that will guide researchers in the reproducibility of their datasets. This decision tree will serve as the basis for a future application that automates the process of data reproduction by automatically supplying the relevant metadata based on specific circumstances. We will develop a web-based service based on such a decision tree to undertake a large portion of the work involved in making our data FAIR (Findable, Accessible, Interoperable, and reusable). In our project, we are primarily concentrating on the SoBigData infrastructure to determine to what extent it adheres to FAIR principles and to delineate the main FAIR issues the decision tree-based FAIR web service must guide to solve.

KEYWORDS

Data Reproducibility, FAIR Principle, Web-based Service, Decision Tree.

ACM Reference Format:
Payel Patra. 2018. Improving SoBigData platform's FAIR principles through decision tree and web-based services. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 CONTEXT

The context of our study is SoBigData [3], the European Big Data and Social Mining Research Infrastructure. It strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData RI will push the FAIR (Findable, Accessible, Interoperable) and FACT (Fair, Accountable, Confidential, and Transparent) principles. It will also orient resources from multiple perspectives: e-infrastructures and online services developers; big data analytics and AI; complex systems focussed on modeling social phenomena; ELSEC (Ethical, Legal, SocioEconomic and Cultural) aspects of data protection; privacy-preserving techniques.

So, because SoBigData platform is an open science platform, handling large amounts of data, it must adhere to the FAIR guidelines [5].

2 PROBLEM

We want to analyze to what extent the SoBigData platform well implements the FAIR principles [5] to highlight potential limitations and issues and to suggest important improvements. To this aim, we first compare it with respect to the other two key repositories: Zenodo [2] and NCBI [4]is a general-purpose open repository created by CERN as part of the European OpenAIRE program. Researchers can store study papers, data sets, research software, reports, and any other digital artifacts linked to their research. The National Center for Biotechnology Information (NCBI) [4] is an online repository for biological information and data that includes the GenBank nucleic acid sequence database and the PubMed database of citations and abstracts published in life science journal articles.

FAIR principles should be considered for any repository that stores massive volumes of data on a daily basis.We report in Table 1 the main findings of our study that highlight that significant concerns continue, as summarized in the table below. The problems with the SoBigData repository are:

- FINDABILITY: This table clearly shows that findable SoBigdata is not responding due to a large number of private files.
- ACCESSABILITY: Access to the maximum amount of articles and datasets is not possible with SoBigData, and users must also log in before uploading and downloading.
- INTEROPERABILITY: some local languages are employed in the interoperability SoBigData platform.Furthermore, relatively few inoperable keywords are used in the abstract section.
- REUSABILITY: Data for reusable, further articles should be publicly available and data reusing is difficult for the SoBigData platform since authors must make their content public.

3 SOLUTION

To improve the implementation of FAIR principles of online platforms and especially of SoBigData, we propose a web-based service

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Fair Principles	SoBigData	NCBI	Zenodo
1. Findable	PID is URL and all articles 70% are private and others are public.	PID is DOI,OMID,PCMID. Maximum articles are public and less are private.	DOI is used as PID. All most all data are public. very few are private.
2. Accessible	Meta data is accessible but almost all data as articles or papers are not accessible. Users must authorize before login to access data.	With PMID and DOI data is accessible .Without account user can download and access the data.	With DOI metadata and data both are accessible for most numbers. MDPI account is openly accessed. So no need to log in.
3. Interoperable	English and Italian both language is used here.References are not given in private mode but in public references are mentioned properly.	English is used only. In NCBI references are there for dataset in PubMed. If it has free PMC portal then references are available.	English is mostly used. In MDPI platform all papers and articles have references to get the information.
4. Findable	Articles are registered but huge steps to get data for reuse purpose as in private mode. And materials under domain are provenance.	The papers are registered with PMCID number and PMID number. here yes, the materials in each resource are provenance.	The papers are registered with a license in the MDPI domain . Yes, in one domain MDPI all datasets and papers are provenances.

Table 1: Comparison of three repositories [i.e SoBigData, NCBI, Zenodo] based on FAIR Principles

shown in Figure 1. The service will leverage on the decision tree published in [1].

The envisaged service (namelly, FARI service) has the aim of forcing scientists and authors to provide needed information (such as raw data, metadata, references, keywords, codes, and so on) about scientific data and datasets. In fact, the decision tree methodology, dealing with all of this information, will assure that FAIR standards are met.

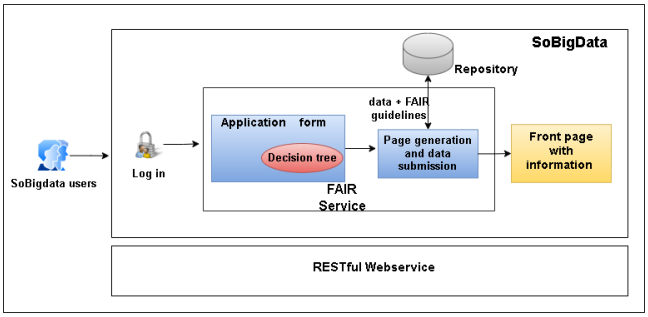

Figure 1: The approach to solving the FAIR principles for SoBigData repository

Figure 1 sketches the workflow we have in mind for the FAIR Service. A SoBigData user logs in to the platform and asks for uploading a new dataset. The platform will run the FAIR service activating the Application form task that is based on the decision tree defined in [1]. The application form, with the decision tree on the backend, will ask questions about the quality and nature of the data, shepherding the user towards Fair standards compliance. The service then submits the dataset in the repository and generates the front page related to the uploaded dataset. If the dataset is not fully compliant with the FAIR principles, the service will generate also FAIR guidelines for the dataset, which are then stored in the repository as well as the dataset and its relative meta-data. The FAIR Service will be implemented using RESTful Webservices and API.

ACKNOWLEDGMENTS

Payel Patra is supported by European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021. I'd like to express my gratitude to Professor Antinisca Di Marco and my supervisor Daniele Di Pompeo, for their amazing guidance, assistance, and support throughout this brilliant and excellent research work.

REFERENCES

[1] Andrea Bianchi, Giordano d'Aloisio, Francesca Marzi, and Antinisca Di Marco. 2023. A Decision Tree to Shepherd Scientists through Data Retrievability. *arXiv preprint arXiv:2304.05767* (2023).

[2] Mathias Dillen, Quentin Groom, Donat Agosti, and Lars Holm Nielsen. 2019. Zenodo, an Archive and Publishing Repository: A tale of two herbarium specimen pilot projects. *Biodiversity Information Science and Standards* 2 (2019).

[3] Valerio Grossi, Beatrice Rapisarda, Fosca Giannotti, and Dino Pedreschi. 2018. Data science at SoBigData: the European research infrastructure for social mining and big data analytics. *International Journal of Data Science and Analytics* 6 (2018), 205–216.

[4] NCBI. [n. d.]. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/. [Online].

[5] 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. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.