MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230946

How to Assess FAIRness of Your Data – A Summary of Testing Two FAIR Validators

Caroline Stellmach[a] and Michael Rusongoza Muzoora[a,1]
[a] *Berlin Institute of Health at Charité- Universitätsmedizin Berlin*
ORCiD ID: Caroline Stellmach https://orcid.org/0000-0001-6798-8533,
Michael Rusongoza Muzoora https://orcid.org/0000-0002-1384-1509

Abstract. Decision-making in healthcare is heavily reliant on data that is findable, accessible, interoperable and reusable (FAIR). Evolving advancements in genomics also heavily rely on FAIR data to steer reliable research for the future. For practical purposes, ensuring FAIRness of a clinical data set can be challenging but could be aided by using FAIR validators. The study describes the test of two open-access web-tools in their demo versions to determine the FAIR levels of three submitted genomic data files with different formats (JSON, TXT, CSV). The F-UJI tool and FAIR-Checker tools provided similar FAIR scores for the three submitted files. However, the F-UJI tool assigned a total rating whereas the FAIR-Checker gave scores clustered by FAIR principles. Neither tool was suited to determine FAIR levels of a FHIR® JSON metadata file. Despite their early developmental status, FAIR validator tools have great potential to assist clinicians in the FAIRification of their research data.

Keywords. FAIR principles, FAIR data, genomic data

1. Introduction

Prevention, diagnosis and treatment of health conditions require data – from clinical exams, laboratory tests as well as from research. Healthcare data also increasingly details genomic information detected by sequencing a patient's biological sample.

The value of this health data depends on characteristics of data described by the FAIR principles; data must be findable, accessible, interoperable and reusable [1].

FAIRness of data is an objective stipulated on an institutional level by the European Commission [2]. The proposed European Health Data Space will be a significant use case for the development and exchange of FAIR healthcare data. On a practical level, clinicians and researchers need validation tools to assess the FAIRness of their data. This study describes our experiences in using the demo versions of two web-based tools, F-UJI and the FAIR-Checker, to assess the FAIRness of three publicly available open-access data files that comprised genomic data, present in different file formats (JSON, TXT, CSV).

[1] Corresponding Author: Michael Rusongoza Muzoora, email: michael.muzoora@bih-charite.de.

2. Methods

A literature search in order to identify available FAIR data validation tools led us to the FAIRassist website which is being developed to provide guidance on the available resources for FAIR data sharing [3] and has identified 19 available resources to assess FAIR data. Out of the 19 resources, we selected two tools which provided demo versions and documentation for testing preselected data.

The FAIRness of the preselected three data files was manually assessed before submitting them for testing. On November 24th of 2022, we submitted the URLs of:

- the 'Genetic variant assessment' BRAF example (FHIR® JSON file, developed by the German Medical Informatics Initiative, short MII) [4],
- SNP in the BRCA gene example (TXT file, NCI GDC Data Portal) [5] and
- the METABRIC dataset (CSV file, available on Kaggle) [6]

one after the other to the automated FAIR data assessment tool F-UJI [7] and the FAIR-Checker [8].

The Genetic variant assessment example was developed by partners in the MII to create an interoperable, FHIR®-based genomics report for Germany. The JSON-file is an example Observation resource, providing details on a detected variant in the BRAF gene, coded using standard terminologies. We expected this file to be assessed by the tools as meeting all four FAIR principles. Secondly, we chose to test a TXT-file on single nucleotide polymorphisms (SNP) in the BRCA gene, publicly available on the National Cancer Institute's (NCI) GDC Data Portal which we expected to also meet the FAIR principles of accessibility and findability but score lower on interoperability due to a lack of syntactical and semantic standards used. The third test was performed using the Breast Cancer Gene Expression Profiles (METABRIC) CSV-file that we believed would show intermediate FAIRness.

3. Results

Table 1 summarizes the key technical features of both FAIR tools which were available as automated web services.

Table 1. Overview of the features of the two tested FAIR validation tools.

F-UJI	FAIR-Checker
Aim: Evaluate FAIRness of research data objects contained in data sets	Aim: Check FAIRness of web resources (web pages)
How it works:	How it works:
For each FAIR principle, the FAIRSFAIR consortium defined one or more metrics and practical tests to evaluate a data set against. These metrics make up a hierarchical model used for the assessment which was created in an iterative process with feedback loops.	1. Extraction of semantic annotations from web page to form 'minimal knowledge graph' 2. Datacite, OpenAire and WikiData used to complete the graph 3. Check to see if Linked Open Vocabularies, Ontology Lookup Services or Bioportal recognize graph's properties and classes
The tool is available as a web demo, R client package and open-source web client [7].	4. Validation of web page metadata against Bioschema's community profiles [8]

3.1. F-UJI

The FAIR assessment performed based on the submission of all three data files' URLs by F-UJI led to the following results:

- Genetic variant assessment example (JSON) – 14%

For the genetic variant assessment example file, the F-UJI tool calculated a FAIR level of 'initial'; scoring highest in terms of findability with 2.5/7, 1/3 in terms of accessibility, interoperability scored 0/4, deeming it to be 'incomplete', and 'initial' for reusability, scoring 0/10.

- SNP example (TXT) - 4% FAIR

F-UJI assessed the FAIR level of the SNP example file as 'initial', providing a 1/7 score for findability, 0/3 for accessibility, 0/4 for interoperability and 0/10 for reusability. The tool recognized that the data file was assigned a globally unique identifier, however no other characteristics were detected.

- METABRIC file (CSV) - 56% FAIR

The METABRIC file received a FAIR level of 'moderate' in terms of findability (score: 3.5/7). Accessibility and interoperability were assessed as 'advanced', scoring 3/3 and 3/4 respectively. Reusability scored 4/10, leading to a level of 'initial'.

The reasoning provided for calculating 'initial' FAIR scores for the first two tested files were, among others, due to not detecting: persistent data content identifiers, RDF-compliant structured metadata, license information and links to semantic vocabularies.

3.2. FAIR-Checker

The results of the assessment performed using the FAIR-Checker are shown in Table 2.

Table 2. Results of the assessment performed on three example files using the FAIR-Checker tool.

Assessment criteria	Genetic variant assessment example (JSON)	SNP example (TXT)	METABRIC file (CSV)
Findability	25%	25%	100%
Principle F1A	success	success	success
F1B	failure	failure	success
F2A	failure	failure	success
F2B	failure	failure	success
Accessibility	100%	100%	100%
Principle A1.1	success	success	success
Interoperability	0%	0%	100%
Principle I1	failure	failure	success
I2	failure	failure	success
I3	failure	failure	success
Reusability	0%	0%	66.77%
Principle R1.1	failure	failure	success
R1.2	failure	failure	success
R1.3	failure	failure	failure

The main reasons provided for a 0% score for interoperability and reusability was that the FAIR-Checker did not detect persistent identifiers, RDF-compliant metadata, as well as license and provenance information in the submitted files.

3.3. User experience

Both tools were simple to use and only required the posting of the example data's URL for assessment. A definition of the expected or acceptable format(s) (tabular, structured or text, etc.) of the submitted data was not provided. While in their usability and performance very similar, the FAIR-Checker tool does not calculate an overall score.

4. Discussion

F-UJI rated the Genetic variant assessment example file as low ('initial') in level of FAIRness contrary to our expectations. Consequently, in its current version, the tool should not be used to determine the level of FAIRness of FHIR resources. The SNP file was also rated 'initial' in FAIRness, with a score of 4% which was lower than expected but more closely matched our expectations. The third file received the highest rating within the test, it was deemed moderately FAIR. Overall, the F-UJI tool was able to detect characteristics specific to all four FAIRness principles in the submitted files.

Like the F-UJI tool, the FAIR-Checker assigned a lower rating of FAIRness to the Genetic variant assessment and the SNP example files and the METABRIC file received a 100% score in three out of four categories. The tests highlight that using RDF-compliant metadata as well as providing license and provenance information within the metadata is crucial to ensure data is reusable and interoperable.

The results of the FAIR tests we performed with three selected files using the F-UJI and FAIR-Checker tools are only indicative of the true potential of these tools since both are still in development and demo versions were used.

It would be useful if the developers provided guidance on the expected data formats on the tool's submission page for both tools tested. Also, facilitating FAIRness of genomic data could be assisted if the tools were compatible with genomic file formats (e.g., VCF, RDF) and able to conduct FAIR assessments for those. Especially since the Global Alliance for Genomics and Health (GA4GH)'s phenopackets schema is being further developed to improve FAIRness of genomic data [11].

5. Conclusions

Our test of two FAIR validation tools using three genomic data files led to three key insights: firstly, the output of the tools differed and was in both cases dependent on the submitted file format. Only one of the tools provided an aggregate FAIR score, although both tools recorded scores for each FAIR principle. Secondly, the tools are not apt for testing the FAIRness of metadata such as FHIR JSON files. Lastly, to facilitate the integration of genomic data into healthcare decision-making, FAIR tools need to be compatible with genomic data specific file formats.

We expect that the availability of reliable FAIR assessment tools and guidelines for establishing FAIR data in their production version will be crucial to support international initiatives such as the development of the European Health Data Space (EHDS). These resources facilitate the fast exchange of health data, including genomic data to improve patient care.

Acknowledgements

We would like to thank the developers of the F-UJI and FAIR-Checker tools for creating these useful resources.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018, doi: 10.1038/sdata.2016.18.

[2] European Commission and Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data. Publications Office. 2018, doi: 10.2777/1524.

[3] FAIRassist and University of Oxford, 'FAIRassist.org'. 2019 Jul 02. https://fairassist.org/#!/ (accessed Nov. 14, 2022).

[4] Medizininformatik Initiative. Genetic variant assessment. 2022. https://simplifier.net/medizininformatikinitiative-modulomics/example-mii-molgen-variante-1 (accessed Nov. 26, 2022).

[5] National Cancer Institute. GOOFS_p_TCGA_b117_118_SNP_N_GenomeWideSNP_6_E06_778014.grch38.seg.v2.txt. 2018 Aug 23. https://portal.gdc.cancer.gov/files/a67a1fe3-e5cc-425f-bb22-e2e7508ffd36 (accessed Nov. 26, 2022).

[6] Breast Cancer Gene Expression Profiles (METABRIC), kaggle, 2019. https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric (accessed Nov. 26, 2022).

[7] Anusuriya Devaraju and Robert Huber. F-UJI - An Automated FAIR Data Assessment Tool (v1.0.0). Zenodo. 2020. https://doi.org/10.5281/zenodo.4063720.

[8] Rosnet, Thomas, Lefort, Vincent, Devignes, Marie-Dominique, and Gaignard, Alban. FAIR-Checker, a web tool to support the findability and reusability of digital life science resources. 2021 Jul, doi: 10.5281/ZENODO.5914307.

[9] Robert Huber *et al.*, 'FAIRsFAIR Data Object Assessment Metrics_v0.4_PublicFeedback', *Google Docs*, Oct. 12, 2020. https://docs.google.com/document/d/1ymkzVmF_BJmKTQZO0SRQ1YQJaPxefIJZ84AKUJUlGeM/edit?usp=sharing&usp=embed_facebook (accessed Nov. 17, 2022).

[10] Fostering FAIR Data Practices In Europe. FAIRsFAIR Data Object Assessment Metrics: Request for comments. FAIRsFAIR. Jul 2020. https://www.fairsfair.eu/fairsfair-data-object-assessment-metrics-request-comments (accessed Nov. 17, 2022).

[11] Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, Callahan TJ, Chute CG, Courtot M, Danis D, Elemento O, Essenwanger A, Freimuth RR, Gargano MA, Groza T, Hamosh A, Harris NL, Kaliyaperumal R, Lloyd KCK, Khalifa A, Krawitz PM, Köhler S, Laraway BJ, Lehväslaiho H, Matalonga L, McMurry JA, Metke-Jimenez A, Mungall CJ, Munoz-Torres MC, Ogishima S, Papakonstantinou A, Piscia D, Pontikos N, Queralt-Rosinach N, Roos M, Sass J, Schofield PN, Seelow D, Siapos A, Smedley D, Smith LD, Steinhaus R, Sundaramurthi JC, Swietlik EM, Thun S, Vasilevsky NA, Wagner AH, Warner JL, Weiland C; GAGH Phenopacket Modeling Consortium; Haendel MA, Robinson PN. The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat Biotechnol. 2022 Jun;40(6):817-820, doi: 10.1038/s41587-022-01357-4.