



@berlinnovation

How to assess FAIRness of your data – a summary of testing two FAIR validators

Michael Rusongoza Muzoora

Research Associate

Berlin Institute of Health at Charité





Agenda

- FAIR Health Data
- Methods
- Results
- Discussion
- Conclusion



Disclosure

- We do not have a conflict of interest with any of the mentioned services provided by F-UJI Tool and FAIR-Checker
- F-UJI Tool and FAIR-Checker were contacted and are aware of our paper



FAIR Health Data

- **Quality data** (exams, laboratory tests, research) needed for prevention, diagnosis and treatment of diseases
- To **extract value from data**, it must be:
- Why is FAIR data important?
 - Objective of FAIR healthcare data (European Commission)
 - Exchange of FAIR data within European Health Data Space (EHDS)
- How FAIR is my data?
 - Need for FAIR data validation tools





FAIR Principles

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

Source: Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3. 160018. <https://doi.org/10.1038/sdata.2016.18>



F-UJI – An Automated FAIR Data Assessment

- Developed by the FAIRsFAIR project, launched in 03/2019

F-UJI Tool Web-based RESTful service; FAIRness assessment based on metrics (17)

- Source code available on GitHub (free license)

Metrics	
FsF-F1-01D	Data is assigned a globally unique identifier
FsF-F1-02D	Data is assigned a persistent identifier
FsF-F2-01M	Metadata includes descriptive core elements to support data findability.
FsF-F3-01M	Metadata includes the identifier of the data it describes.
FsF-F4-01M	Metadata is offered in such a way that it can be retrieved by machines.
FsF-A1-01M	Metadata contains access level and access conditions of the data.
FsF-A1-02M	Metadata is accessible through a standardized communication protocol
FsF-A1-03D	Data is accessible through a standardized communication protocol
FsF-A2-01M	Metadata remains available, even if the data is no longer available.
FsF-I1-01M	Metadata is represented using a formal knowledge representation language.
FsF-I1-02M	Metadata uses semantic resources.
FsF-I3-01M	Metadata includes links between the data and its related entities.
FsF-R1-01MD	Metadata specifies the content of the data.
FsF-R1.1-01M	Metadata includes license information under which data can be reused.
FsF-R1.2-01M	Metadata includes provenance information about data creation or generation.
FsF-R1.3-01M	Metadata follows a standard recommended by the target research community of the data.
FsF-R1.3-02D	Data is available in a file format recommended by the target research community.

Modified from: <https://www.f-uji.net/index.php?action=methods>



FAIR-Checker

- Developed by the Interoperability working group at French Institute of Bioinformatics
- Web-based tool; 4-step process
 - Semantic web technologies used to check for use of standards

FAIR-Checker

1. Knowledge graph build from semantic annotations (in metadata)
2. Existent knowledge graphs used to complete graph
3. Test if classes/properties in graph are recognized using DLS/LOV/Bioportal
4. Validation of metadata against Bioschemas community profiles

Principle	Name
F1A	Unique IDs
F1B	Persistent IDs
F2A	Structured metadata
F2B	Shared vocabularies for metadata
A1.1	Open resolution protocol
I1	Machine readable format
I2	Use shared ontologies
I3	External links
R1.1	Metadata includes license
R1.2	Metadata includes provenance
R1.3	Community standards



Methods

1. Identify available **FAIR data validation tools** through literature search



[FAIRassist website](#)

2. Select two validation tools for **testing**



[F-UJI – Automated FAIR Data Assessment Tool](#)
&
[FAIR-Checker](#)

3. Test both tools **using three variant data files** with different formats (JSON, TXT, CSV)



Methods – Resources tested (1/3)

1. Example FHIR® Observation resource of a detected genetic variant

```
{
  "resourceType": "Observation",
  "id": "mi-exa-molgen-variante-1",
  "meta": {
    "profile": [
      "https://www.medinformatik-initiative.de/fhir/ext/modul-molgen/StructureDefinition/variante",
      "http://hl7.org/fhir/uv/genomics-reporting/StructureDefinition/variant"
    ]
  },
  "code": {
    "coding": [
      {
        "system": "http://loinc.org",
        "code": "69548-6",
        "display": "Genetic variant assessment"
      }
    ]
  },
  "category": [
    {
      "coding": [
        {
          "system": "http://terminology.hl7.org/CodeSystem/observation-category",
          "code": "laboratory",
          "display": "Laboratory"
        }
      ]
    }
  ],
  "component": [
    {
      "code": {
        "coding": [
          {
            "system": "http://loinc.org",
            "code": "48018-6",
            "display": "Gene studied [ID]"
          }
        ]
      }
    }
  ]
}
```



- Fast Interoperability Healthcare Resources (FHIR®)
- Standard exchange format (JSON/XML) for healthcare data
- Structures information into building blocks = „resources“

- Source: German Medical Informatics Initiative (MII)
- Format: JSON



Methods – Resources tested (2/3)

2. Example of a Single Nucleotide Polymorphism (SNP) in the BRCA gene

GDC_Allele	Chromosome	Start	End	Num_Probes	Segment_Mean
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	62920	3298168	737	-0.0858
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	3301765	15547820	7446	-0.1723
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	15551554	16544783	452	-0.1094
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	16569132	16583114	3	-1.7836
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	16604834	16782786	46	-0.364
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	16784801	16788139	5	-1.7637
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	16855942	17010382	82	0.0859
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	17014383	17016429	4	-1.6115
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	17018588	21992508	3311	-0.1641
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	21996664	22001786	2	-2.9341
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	22002025	25230200	1907	-0.1583
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	25232069	25320455	35	0.3834
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	25336819	28740003	1413	-0.1428
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	28753159	34626073	3532	0.2274
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	34637053	34649941	22	-0.41
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	34650960	61661425	16476	0.213
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	61661912	61676223	15	0.6696
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	61676425	65419674	2760	0.2119
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	65419886	65442811	14	0.6801
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	65444442	72284670	4480	0.208
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	72290430	72297687	9	-0.6471
e50f848e-a469-41c1-b44b-4dbc283fd0a6	1	72302735	72345465	48	1.0941

- SNP: genomic variant at a single base position in DNA

- Source: National Cancer Institute, Genomic Data Commons (GDC) Data

Portal

- Project: TCGA-BRCA, Breast Invasive Carcinoma

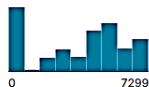
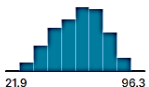
- Format: TXT





Methods – Resources tested (3/3)

3. Breast Cancer Gene Expression Profiles (METABRIC) dataset

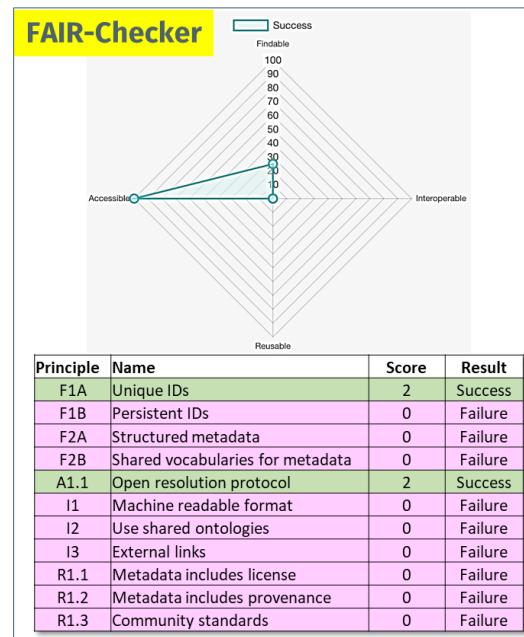
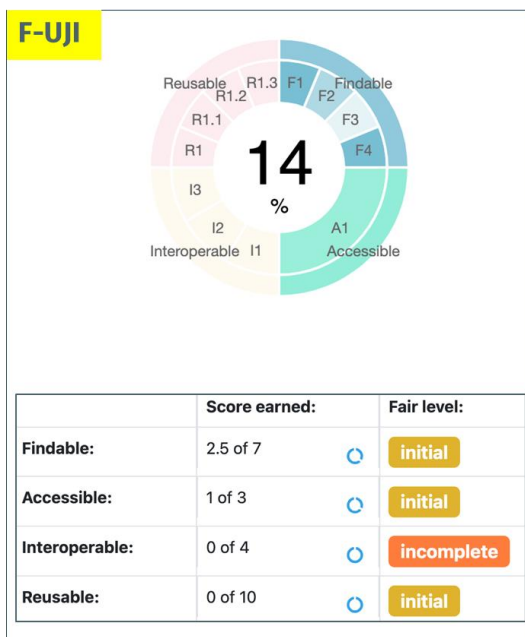
patient_id	# age_at_diagnosis	A type_of_breast_s...	A cancer_type	A cancer_type_deta...	A cellularity	# chemotherapy
		MASTECTOMY 59% BREAST CONSER... 40% Other (22) 1%	Breast Cancer 100% Breast Sarcoma 0%	Breast Invasive Du... 79% Breast Mixed Duct... 11% Other (197) 10%	High 49% Moderate 37% Other (254) 13%	
0	72.99					0
0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma		0
2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0
5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	1
6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	1
8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	1
10	78.77	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0
14	56.45	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	1

- Source: Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)
- Dataset: by Prof. C. Caldas, from [BioPortal](#)
- Format: CSV



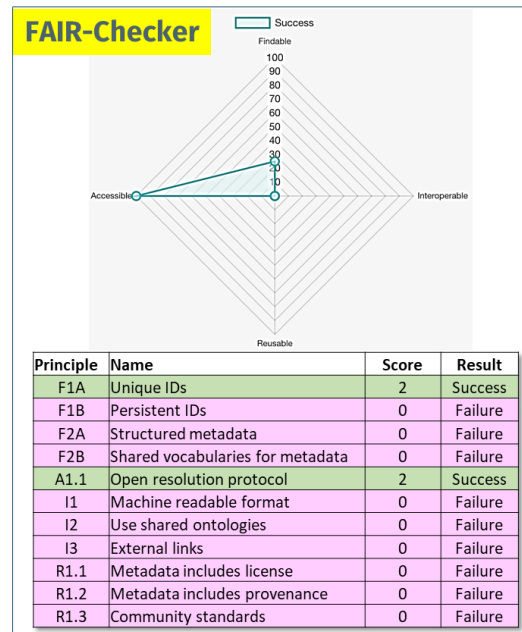
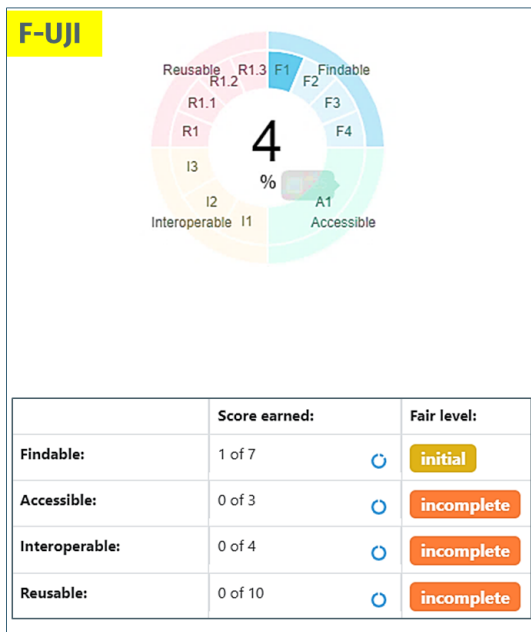


Result 1: Example FHIR® *Observation* resource, MII (JSON)



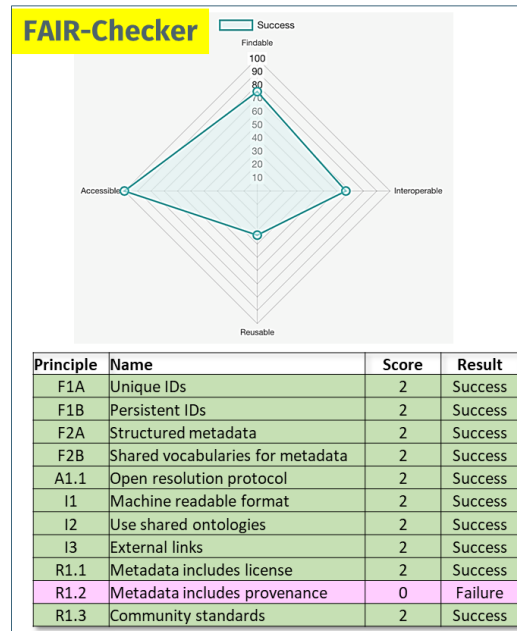
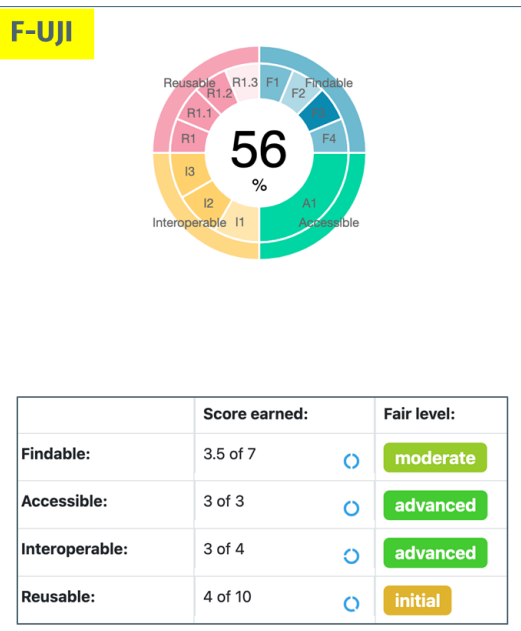


Result 2: Example SNP, NCI (TXT)





Result 3: Gene expression profiles, Kaggle (CSV)





Summary

F-UJI Tool

JSON file: FAIRness low (14%); contrary to expectations

TXT file: FAIRness low (4%); lower than expected

CSV file: moderately FAIR (56%)

→ Tool detected characteristics specific to all 4 FAIR principles in files

→ Current version not suitable for FHIR JSON format



Summary

FAIR-Checker

JSON file: Unique IDs (FIA) & open resolution protocol (AI.I) detected, others failed

TXT file: Unique IDs (FIA) & open resolution protocol (AI.I) detected, others failed

CSV file: Meets findability, accessibility and interoperability principles, but 2/3 reusability principles

→ Tests highlight reusability & interoperability of data can be ensured through:

- Use of RDF-compliant metadata &
- Providing license and provenance information within metadata



Discussion

1. Tests performed with F-UJI and FAIR-Checker are only indicative of the true potential of the tools; they are **still in development** (demo versions were used)
2. It would be useful if **guidance on expected resource format** were provided to users
3. Facilitating FAIRness of genomic data could be assisted if tools were **compatible with genomic file formats** (e.g., VCF)
 - Output of the tools differed and was dependent on the submitted file format.
 - Only one of the tools provided an aggregate FAIR score, both tools recorded scores for each FAIR principle



References

- M. D. Wilkinson et al., 'The FAIR Guiding Principles for scientific data management and stewardship.', Sci. Data, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
- 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.
- FAIRassist and University of Oxford, 'FAIRassist.org', Jul. 02, 2019. <https://fairassist.org/#/> (accessed Nov. 14, 2022).
- Medizininformatik Initiative, 'Genetic variant assessment', 2022. <https://simplifier.net/medizininformatikinitiative-modulomics/example-miimolgen-variante-l> (accessed Nov. 26, 2022).
- National Cancer Institute, 'GDOFS_p_TCGA_b117_118_SNP_N_GenomeWideSNP_6_ED6_778014.grch38.seq.v2.txt', Aug. 23, 2018. <https://portal.gdc.cancer.gov/files/a67af3e3-e5cc-425f-bb22-e2e7508ffd36> (accessed Nov. 26, 2022).
- 'Breast Cancer Gene Expression Profiles (METABRIC)', kaggle, 2019. <https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expressionprofiles-metabric> (accessed Nov. 26, 2022).
- Anusuriya Devaraju and Robert Huber, 'F-UJI - An Automated FAIR Data Assessment Tool (v1.0.0)', Zenodo, 2020. doi: 10.5281/zenodo.4063720
- 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', Jul. 2021, doi: 10.5281/ZENODO.5914307.
- Robert Huber et al., 'FAIRsFAIR Data Object Assessment Metrics_v0.4_PublicFeedback', Google Docs, Oct. 12, 2020. https://docs.google.com/document/d/lymkzVmF_BJmKTQZDQSRQIYQJaPxeFlJZ84AKUJUIGeM/edit?usp=sharing&usp=embed_facebook (accessed Nov. 17, 2022).
- FAIRsFAIR "Fostering FAIR Data Practices In Europe", 'FAIRsFAIR Data Object Assessment Metrics: Request for comments', FAIRsFAIR, Jul. 10, 2020. <https://www.fairsfair.eu/fairsfair-data-object-assessment-metrics-requestcomments> (accessed Nov. 17, 2022).
- J. O. B. Jacobsen et al., 'The GA4GH Phenopacket schema defines a computable representation of clinical data', Nat. Biotechnol., vol. 40, no. 6, Art. no. 6, Jun. 2022, doi: 10.1038/s41587-022-01357-4.

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Thank You!



Michael Rusongoza Muzoora
michael.muzoora@bih-charite.de



Caroline Stellmach
caroline.stellmach@bih-charite.de

