Making and assessing FAIR biomedical data

Medical Informatics Europe (Nice, France)

Chang Sun, Vincent Emonet, Michel Dumontier

Institute of Data Science, Maastricht University, The Netherlands

29-08-2022

Maastricht University

Institute of Data Science

Medical Informatics Europe
MIE2022
EFMI

This tutorial will provide you a clear instruction on

- what **FAIR** is really about
- what steps are needed to **create FAIR biomedical/health data**
- how to **assess the FAIRness** of online digital resources
- how to create and use **domain- specific FAIRness** tests.

Learning Objectives

1. To learn how to make digital resources FAIR and to improve their FAIRness
2. To understand automated FAIRness evaluation and executable metrics
3. To use existing tools to perform FAIRness evaluation
4. To modify or develop a custom collection of evaluation metrics
5. To create and publish new FAIR tests based on domain-specific requirements

Outline

	Time	Topic
Part 1	07:45 - 08:30	Presentation: Welcome + Introduction to FAIR principle, implementation details and FAIR assessment and tools.
	08:30 - 08:40	Break
Part 2	08:40 - 09:30	Hands-on: Assess FAIRness of selected biomedical resources using FAIR assessment tools.
	09:30 - 09:40	Break
Part 3	09:40 - 10:15	Hands-on: Create a custom FAIR metrics test for Biomedical data
	10:15 - 10:30	Interaction: Discussion and closing

PART 1 - Welcome + Intro to FAIR

07:45 - 08:30

- Introduce the FAIR Guiding Principles
- Discuss FAIR data recipes and corresponding implementation details
    - How to make data fair
- Describe FAIR assessment in terms of approaches, metrics, and tools
- Present FAIR evaluation tools and services

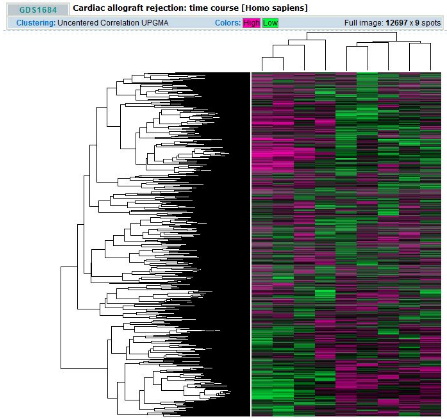
Great opportunities abound with increasing amounts of public and private data

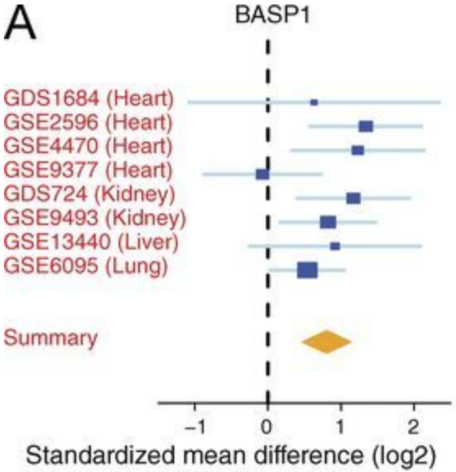
iNSiGHT

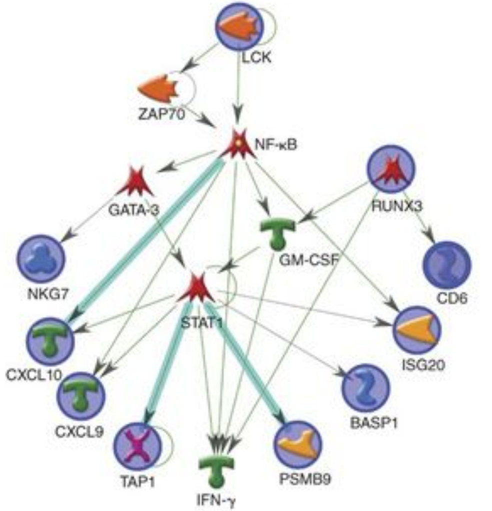
A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation

Khatri et al. JEM. 210 (11): 2205 DOI: 10.1084/jem.20122709

Gene Expression Omnibus







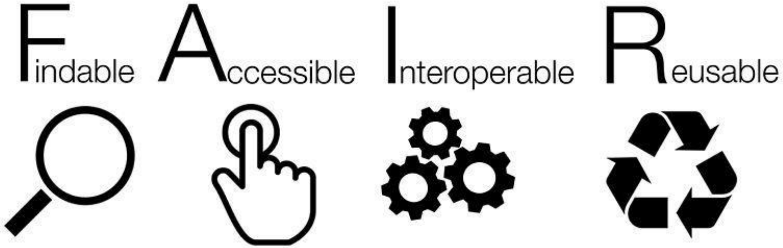
Main Findings:

1. CRM of 11 overexpressed genes **predicted future injury** to a graft
2. Mice treated with **existing drugs against specific CRM genes extended graft survival**
3. Retrospective **EHR data analysis supports treatment prediction**

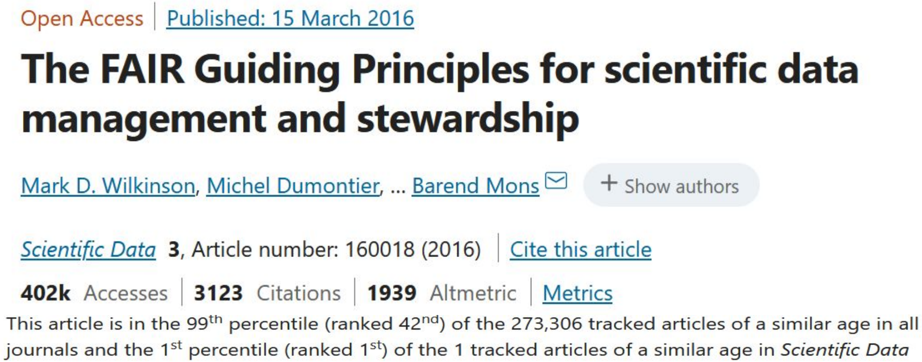
Key Observations:

1. **Meta-analysis** offers a **more reliable estimate** of the direction and magnitude of the effect
2. Existing data can be used to **generate and validate new hypotheses**

However, *significant effort* is still needed to find the right dataset(s), make sense of them, and use for a new purpose



A set of principles to promote the discovery and reuse of digital content *for people and the machines they use*



Open Access | Published: 15 March 2016

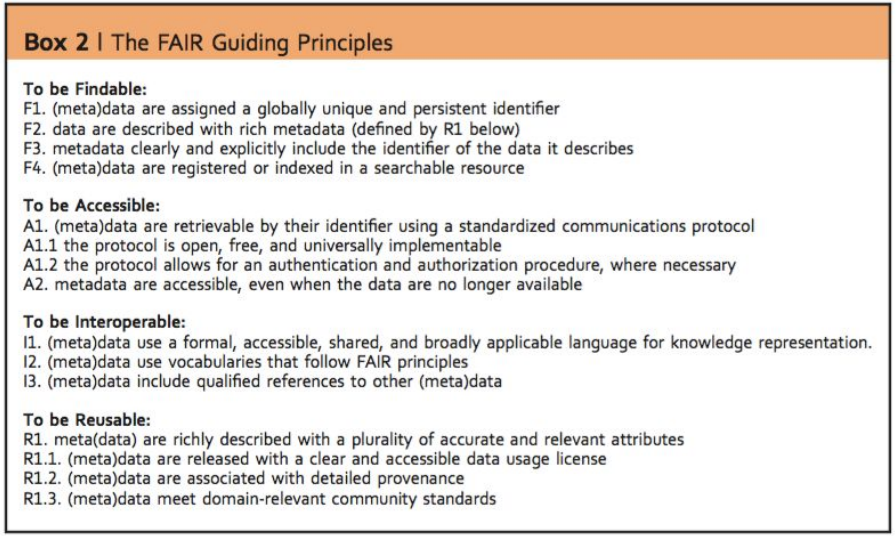
The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, ... Barend Mons ✉ + Show authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

402k Accesses | **3123** Citations | **1939** Altmetric | Metrics

This article is in the 99th percentile (ranked 42nd) of the 273,306 tracked articles of a similar age in all journals and the 1st percentile (ranked 1st) of the 1 tracked articles of a similar age in *Scientific Data*



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

EUROPEAN COMMISSION
Press Release Database

European Commission > Press releases database > Press Release details

European Commission - Statement

G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

1. We, the Leaders of the G20, met in Hangzhou, China on 4-5 September 2016.

G7 2017 ITALIA

Annex 4:
G7 Expert Group on Open Science

Turin, Italy, September 28, 2017

Realising the European Open Science Cloud

First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

Final Report and Action Plan from the European Commission Expert Group on FAIR Data

TURNING FAIR INTO REALITY

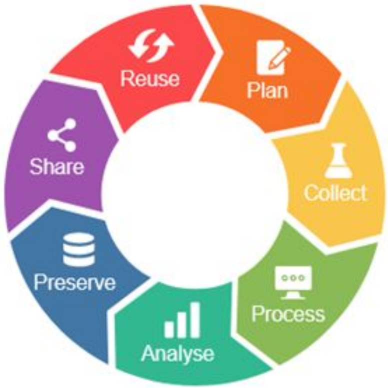
2018

http://www.nature.com/articles/sdata201618

Why should researchers make their data FAIR?

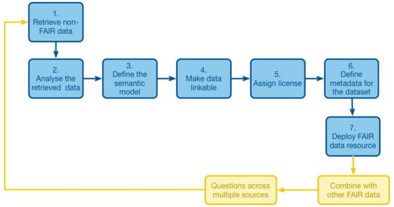
- **Higher impact** of published research results
  - Increased citation owing to reuse/comparison
  - Increased opportunities for collaboration
  - Increased recognition of other outputs as part of performance
- **Reduced communication** in the reproducibility of research results
- **Transfer of responsibility** for the long term preservation of research results
- **Meet the expectations** of funders, institutions, and peers

Learn how to create and publish FAIR data



RDMkit

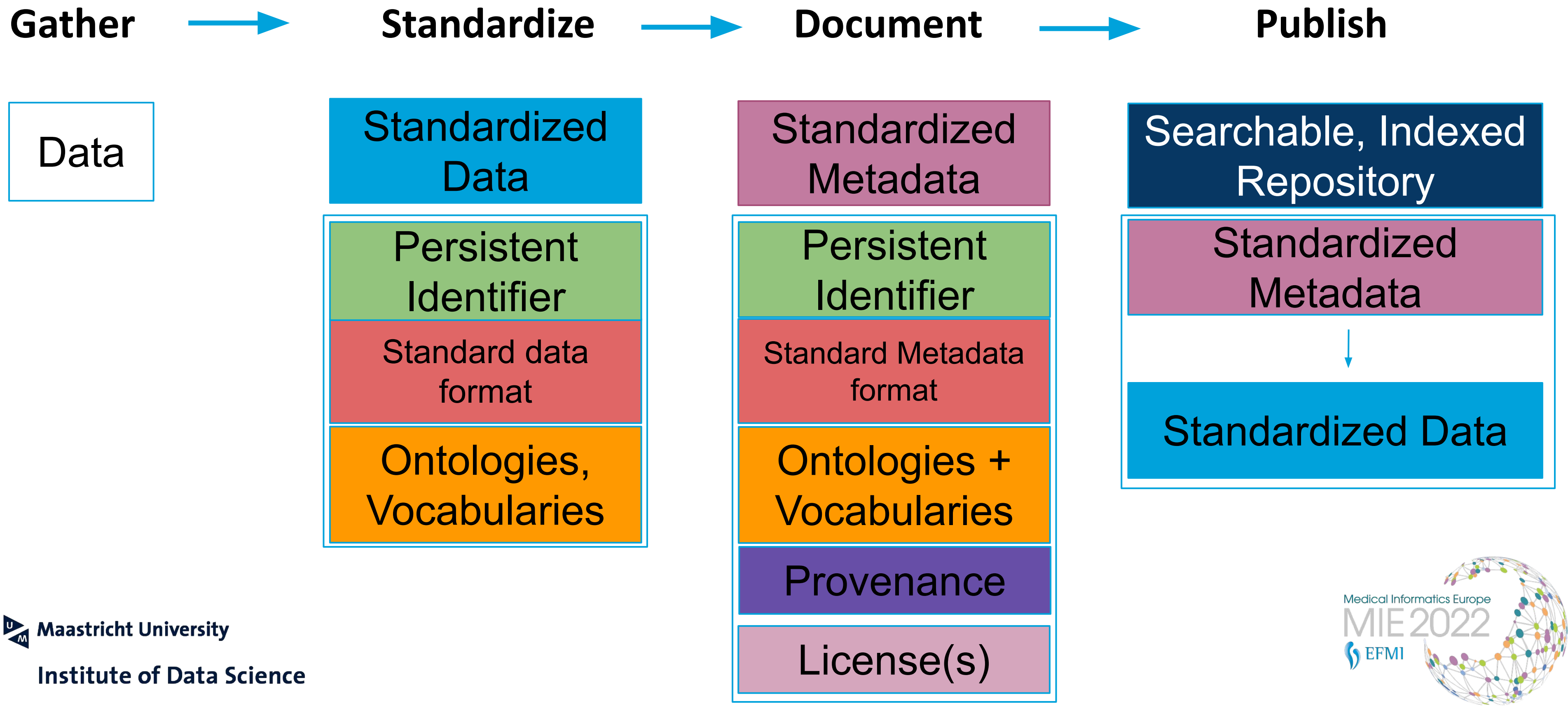
GO FAIR


1. Retrieve non-FAIR data
2. Analyse the retrieved data
3. Define the semantic model
4. Make data linkable
5. Assign license
6. Define metadata for the dataset
7. Deploy FAIR data resource
Combine with other FAIR data
Questions across multiple sources

FAIR cookbook

FAIRification

Findable Accessible Interoperable Reusable



Gather → Standardize → Document → Publish

Data

Standardized Data

Persistent Identifier

Standard data format

Ontologies, Vocabularies

Standardized Metadata

Persistent Identifier

Standard Metadata format

Ontologies + Vocabularies

Provenance

License(s)

Searchable, Indexed Repository

Standardized Metadata

↓

Standardized Data

(meta)data for machines

The long term objective of FAIR is to make content accessible by machines, to support the everyday work we do

RDMkit

(meta)data for machines

- facilitate **query** and **filter** content based on specific variables, experimental conditions, biological sources, and other parameters
- easier to **understand** and **compare** experiments
- easier to **replicate** experiments and **reproduce** research results
- easier to **integrate** data from multiple datasets and studies, sharing the same experimental conditions or variables
- **exchange** content between different tools and environment
- **explore** and **visualize** knowledge connections
- **query** across a number of disparate databases and APIs

(meta)data for machines

The long term objective of FAIR is to make content accessible to machines, to support the everyday work we do

Data and their metadata ought to be:

- machine readable - the syntax of the data are formally specified to enable reliable reading/writing of the data.
- machine interpretable - the semantics of the data elements are well defined and can be reasoned about for information retrieval and query answering

Machine readability

tab-delimited files (spreadsheets) are extremely popular owing to their simplicity and human accessibility.

on their own, it is unclear how many rows or columns there should be, nor what the rows or columns represent, nor what the values should be constrained to (if at all)

18463	32	0	94533
18465	55	1	94532
18468	12	0	94533

Machine readability

adding a column header hints to humans as to what the columns may represent, but this is not always the case, and it is still unclear what the value sets should be.

PATIENT	AGE	S	ADDRESS
18463	32	0	94533
18465	55	1	94532
18468	12	0	94533

Machine readability

more questions emerge on closer examination of the data ... what unit is the age (hours, months or years)? what do the values 0,1 represent? what are the allowable values for these columns?

PATIENT	AGE	S	ADDRESS
18463	32	0	94533
18465	55	1	94532
18468	12	0	94533

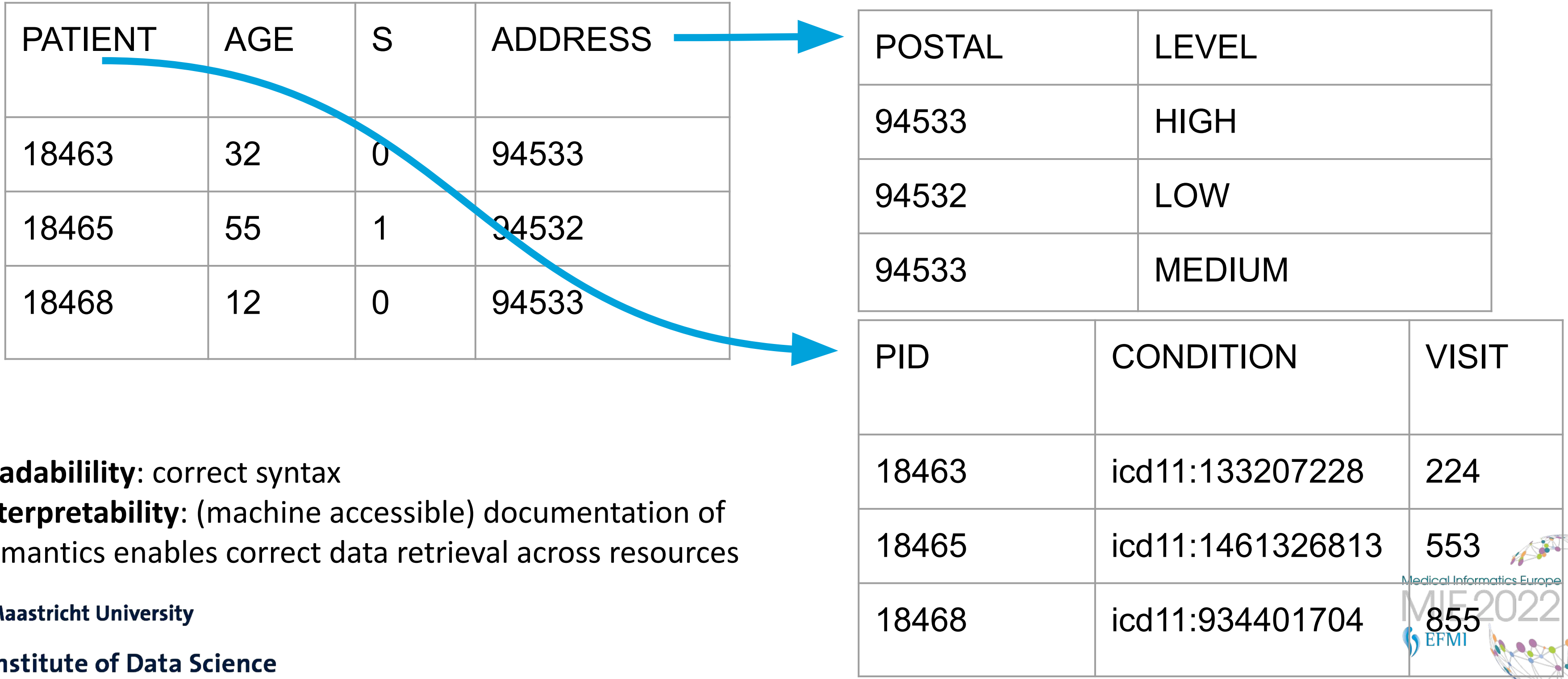
unit?

code book?

format?

Machine intepretability

what is the relative risk of developing respiratory track conditions in areas of high industrial pollution? answering this question requires access to other data tables - there needs to be a correspondance between columns



PATIENT	AGE	S	ADDRESS
18463	32	0	94533
18465	55	1	94532
18468	12	0	94533

POSTAL	LEVEL
94533	HIGH
94532	LOW
94533	MEDIUM

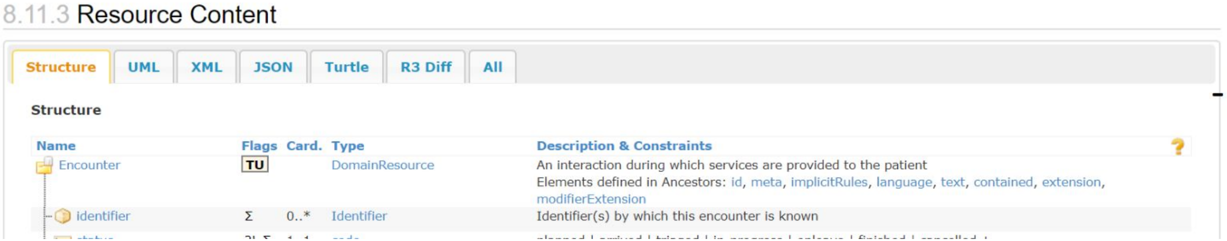
PID	CONDITION	VISIT
18463	icd11:133207228	224
18465	icd11:1461326813	553
18468	icd11:934401704	855

readabililty: correct syntax
interpretability: (machine accessible) documentation of semantics enables correct data retrieval across resources

FHIR

HL7® FHIR® (Fast Healthcare Interoperability Resources) standard for clinical and administrative data.

Free to use, supported by major vendors, foundation in web standards: HTTP, OAuth, formats (JSON, XML, RDF)

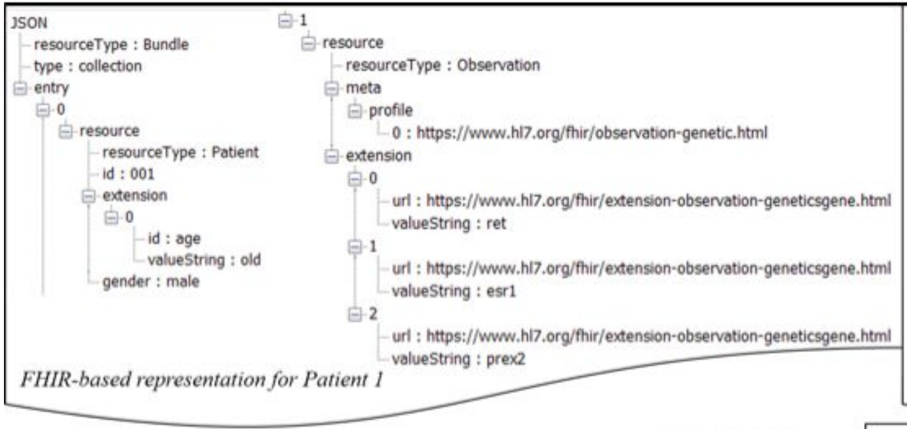


8.11.3 Resource Content

Structure

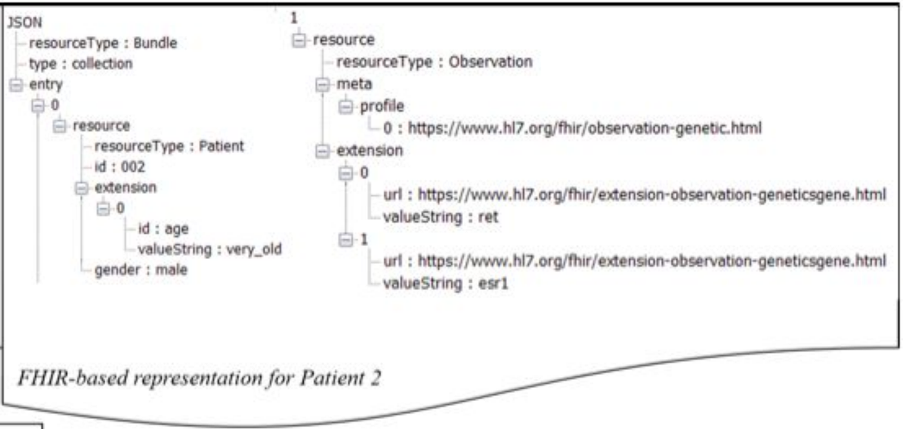
Name	Flags	Card.	Type	Description & Constraints
Encounter	TU		DomainResource	An interaction during which services are provided to the patient Elements defined in Ancestors: id, meta, implicitRules, language, text, contained, extension, modifierExtension
identifier	Σ	0..*	Identifier	Identifier(s) by which this encounter is known

FHIR RDF enables automated inference and integration of clinical and biomedical data models.



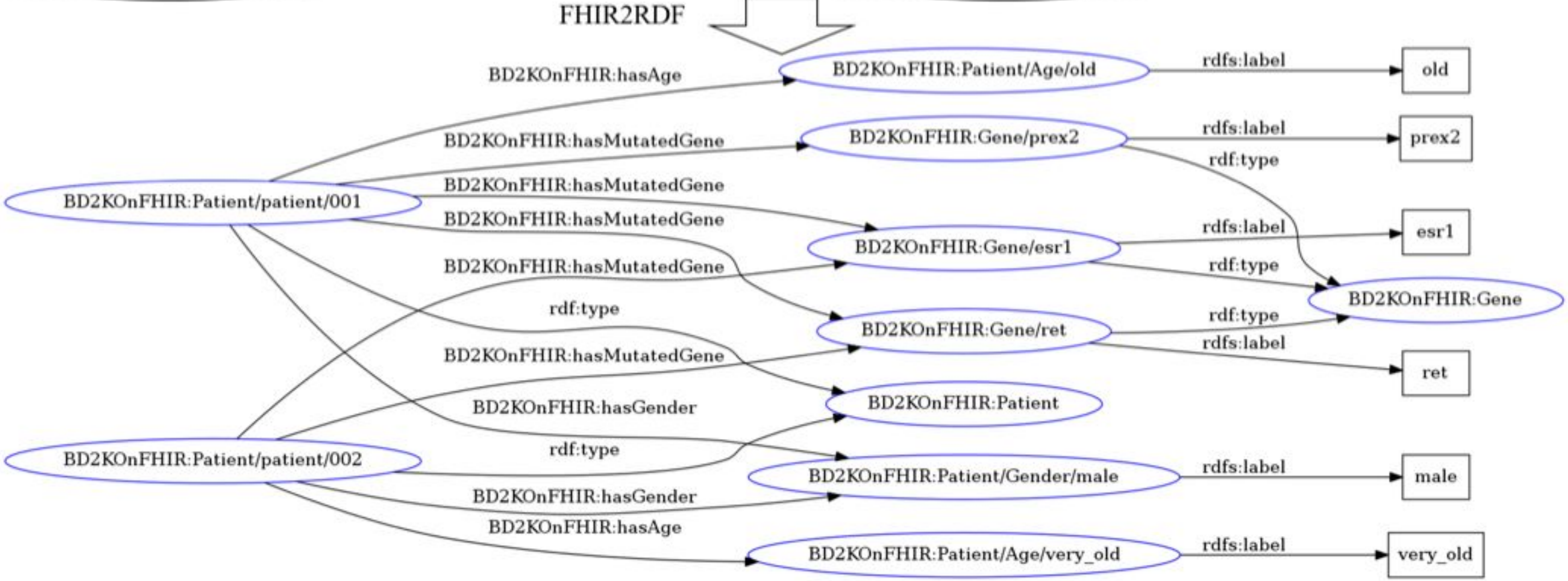
```
JSON
  resourceType : Bundle
  type : collection
  entry
    0
      resource
        resourceType : Patient
        id : 001
        extension
          0
            id : age
            valueString : old
        gender : male
    1
      resource
        resourceType : Observation
        meta
          profile
            0 : https://www.hl7.org/fhir/observation-genetic.html
        extension
          0
            url : https://www.hl7.org/fhir/extension-observation-geneticsgene.html
            valueString : ret
          1
            url : https://www.hl7.org/fhir/extension-observation-geneticsgene.html
            valueString : esr1
          2
            url : https://www.hl7.org/fhir/extension-observation-geneticsgene.html
            valueString : prex2
```

FHIR-based representation for Patient 1



```
JSON
  resourceType : Bundle
  type : collection
  entry
    0
      resource
        resourceType : Patient
        id : 002
        extension
          0
            id : age
            valueString : very_old
        gender : male
    1
      resource
        resourceType : Observation
        meta
          profile
            0 : https://www.hl7.org/fhir/observation-genetic.html
        extension
          0
            url : https://www.hl7.org/fhir/extension-observation-geneticsgene.html
            valueString : ret
          1
            url : https://www.hl7.org/fhir/extension-observation-geneticsgene.html
            valueString : esr1
```

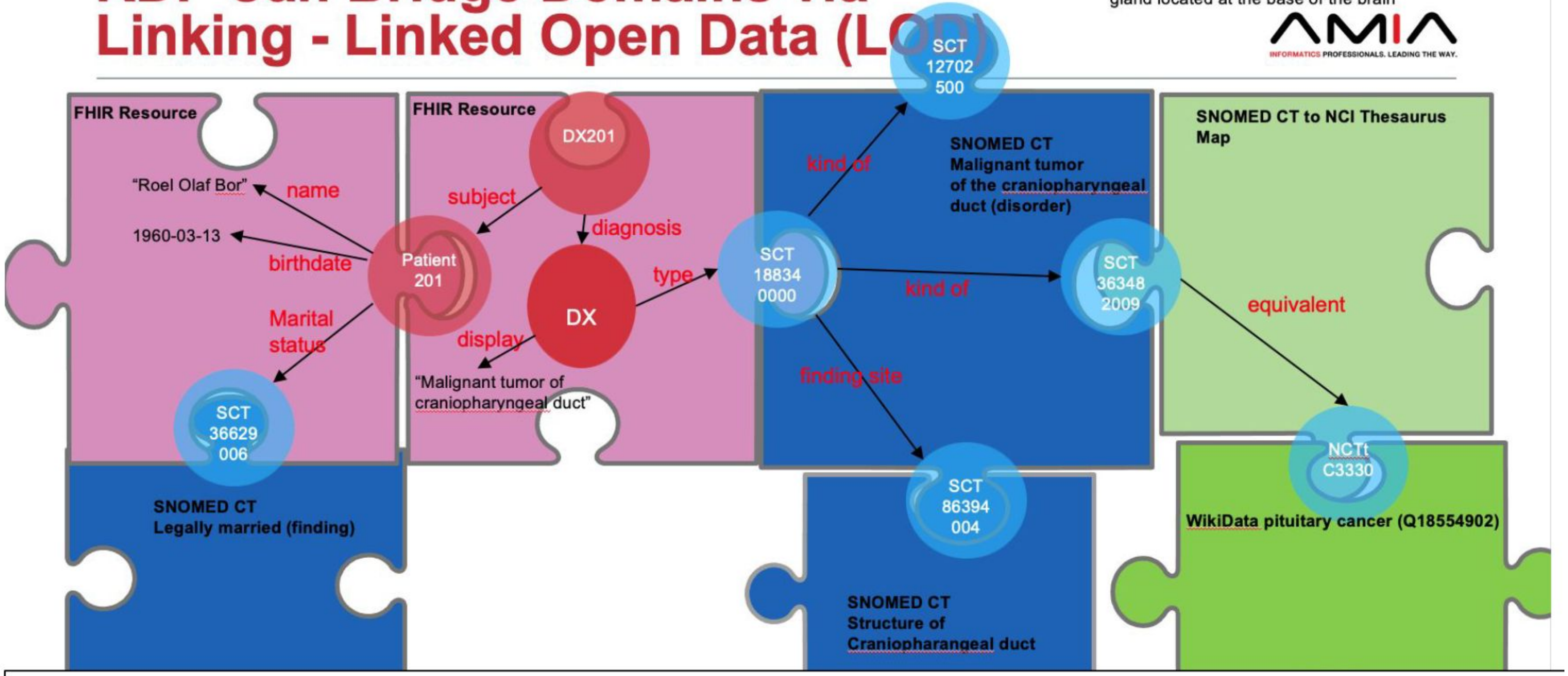
FHIR-based representation for Patient 2



RDF Can Bridge Domains via Linking - Linked Open Data (LOD)

endocrine gland cancer located in the pituitary
gland located at the base of the brain

INFORMATICS PROFESSIONALS. LEADING THE WAY.



Exploring JSON-LD as an Executable Definition of FHIR RDF to Enable Semantics of FHIR Data

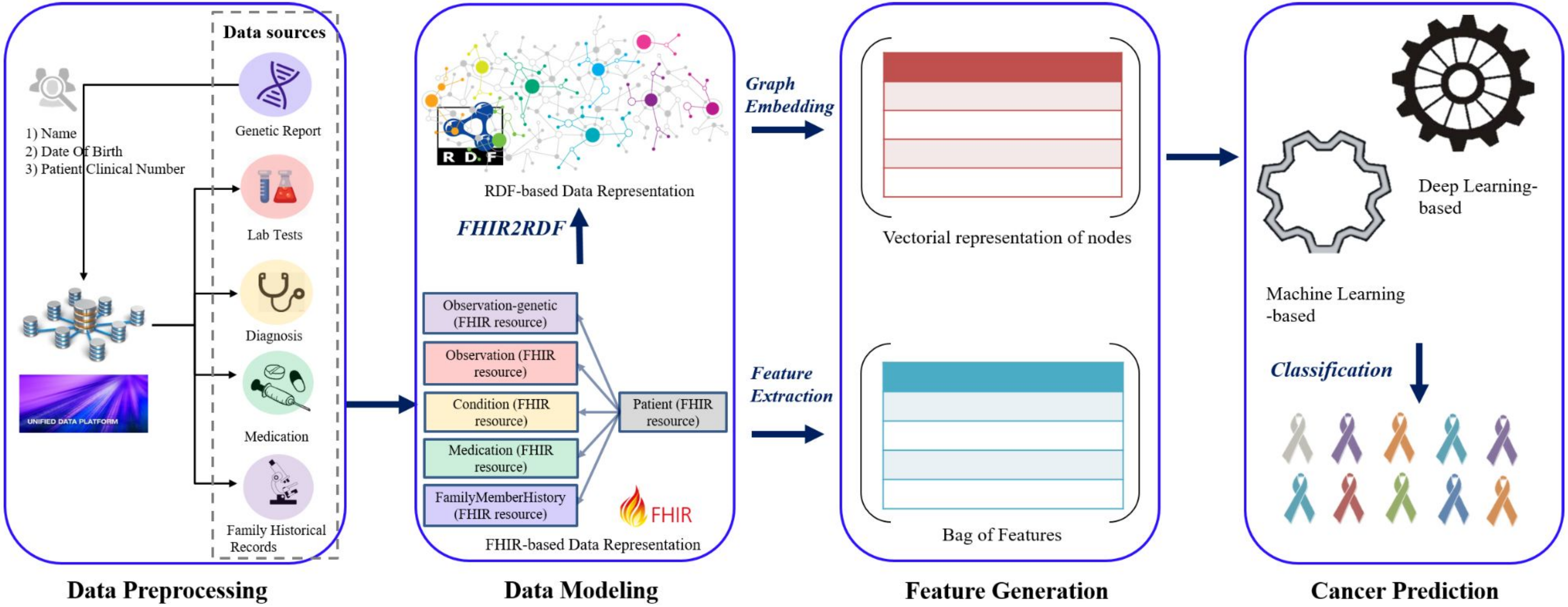
Dazhi Jao[1], Eric Prud'hommeaux[2, 3], David Booth[4], Cory. M Endle[5], Daniel J Stone[5], Guoqian Jiang[5]

Maastricht University
Institute of Data Science

https://yosemiteproject.github.io/Tutorial-FHIR-RDF-as-a-Bridge/

MIE 2022
EFMI

Using FHIR Data for Cancer Research



Maastricht University

Institute of Data Science

https://github.com/fhircat/cancer-prediction-on-fhir-rdf

Medical Informatics Europe MIE2022 EFMI

Phenopackets

Phenopackets is a standard developed by the Global Alliance for Genomics and Health (GA4GH)

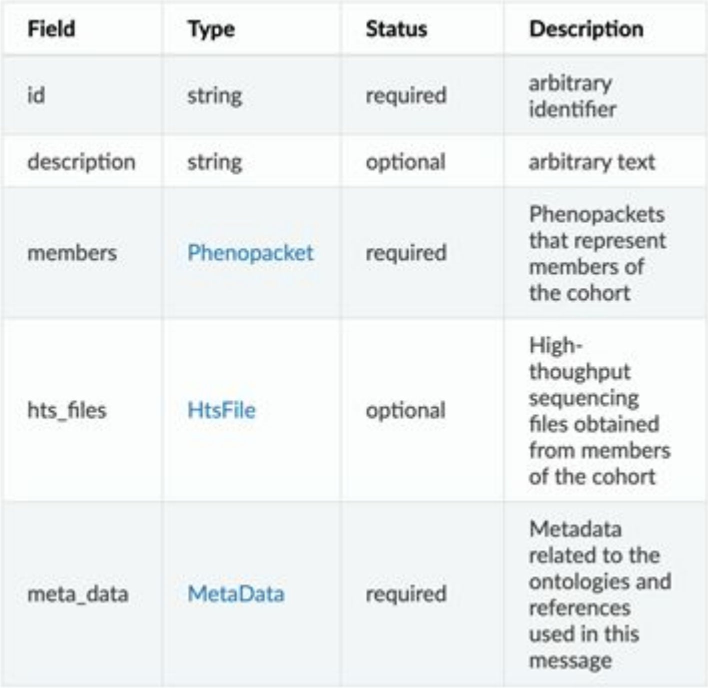
Provides a mechanism for sharing patient phenotype information in a structured and computable manner

Phenopackets specification:

- https://phenopackets-schema.readthedocs.io/en/latest/
- https://github.com/phenopackets

Initiative to make Phenopackets FHIR compatible.

https://phenopacket-schema.readthedocs.io/en/latest/fhir.html



Field	Type	Status	Description
id	string	required	arbitrary identifier
description	string	optional	arbitrary text
members	Phenopacket	required	Phenopackets that represent members of the cohort
hts_files	HtsFile	optional	High-thoughput sequencing files obtained from members of the cohort
meta_data	MetaData	required	Metadata related to the ontologies and references used in this message

Sensitive Data

General Data Protection Regulation (GDPR) addresses personal data about individuals that requires careful consideration. GDPR “**special category data**” prescribes very strict rules involving racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data, data concerning health or data concerning a natural person’s sex life or sexual orientation.

Pseudonymization replaces identifying fields with artificial identifiers, and there is typically a key to reverse identify.

Data anonymization aims to make re-identification of data subjects impossible from these data. Many techniques available including k-anonymity, l-diversity, and differential privacy. Note that it could still be possible to re-identify using other means.

FAIR expects the publication of Metadata that indicates how the data were processed, and how they can be made available to others.

Data Repositories *should* make data more FAIR

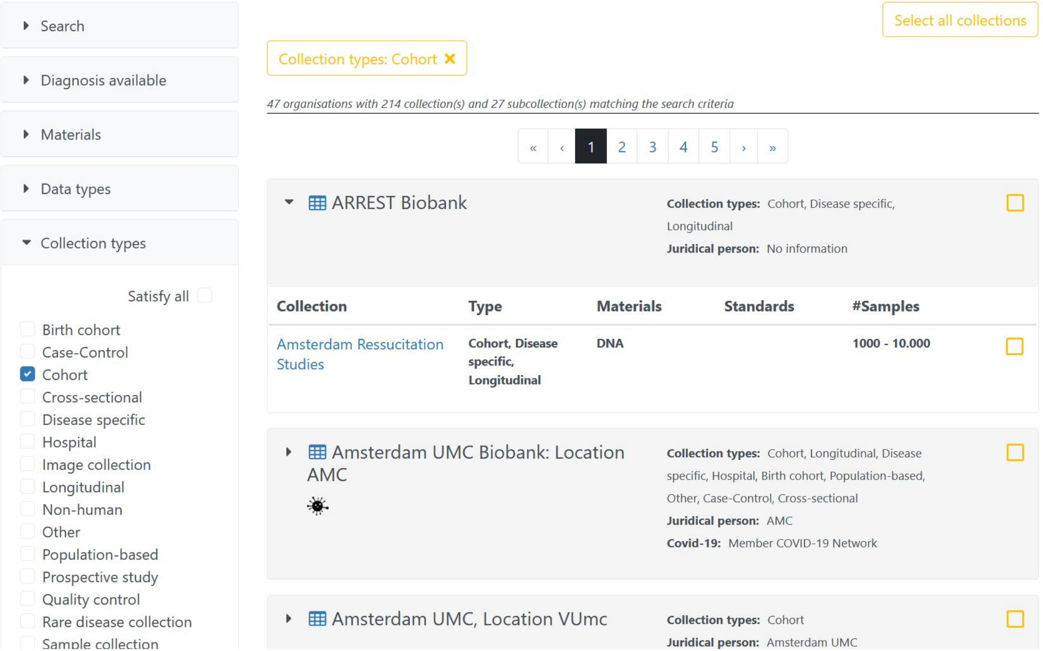
BBMRI-NL

Collections of samples, data, and biobanks in the Netherlands.

Enables ontology-backed metadata description.

Constructed with Molgenis software.

Search and filter for entries of a certain type e.g. cohort studies.



47 organisations with 214 collection(s) and 27 subcollection(s) matching the search criteria

▾ ARREST Biobank □

Collection types: Cohort, Disease specific, Longitudinal

Juridical person: No information

Collection	Type	Materials	Standards	#Samples	
Amsterdam Ressucitation Studies	Cohort, Disease specific, Longitudinal	DNA		1000 - 10.000	□

▸ Amsterdam UMC Biobank: Location AMC □

Collection types: Cohort, Longitudinal, Disease specific, Hospital, Birth cohort, Population-based, Other, Case-Control, Cross-sectional

Juridical person: AMC

Covid-19: Member COVID-19 Network

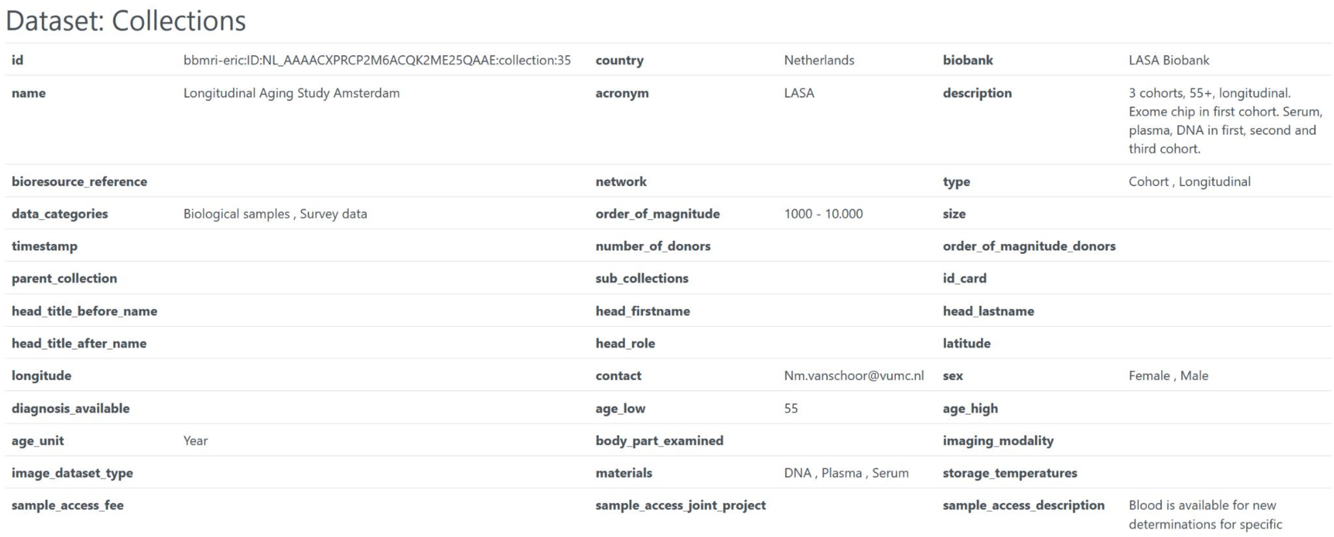
▸ Amsterdam UMC, Location VUmc □

Collection types: Cohort

Juridical person: Amsterdam UMC

Example Entry

Longitudinal Aging Study Amsterdam (LASA) metadata record in BBMRI-NL biobank catalogue V2



Dataset: Collections

id	bbmri-eric:ID:NL_AAAACXPRCP2M6ACQK2ME25QAAE:collection:35	**country**	Netherlands	**biobank**	LASA Biobank
name	Longitudinal Aging Study Amsterdam	**acronym**	LASA	**description**	3 cohorts, 55+, longitudinal. Exome chip in first cohort. Serum, plasma, DNA in first, second and third cohort.
bioresource_reference		**network**		**type**	Cohort , Longitudinal
data_categories	Biological samples , Survey data	**order_of_magnitude**	1000 - 10.000	**size**	
timestamp		**number_of_donors**		**order_of_magnitude_donors**	
parent_collection		**sub_collections**		**id_card**	
head_title_before_name		**head_firstname**		**head_lastname**	
head_title_after_name		**head_role**		**latitude**	
longitude		**contact**	Nm.vanschoor@vumc.nl	**sex**	Female , Male
diagnosis_available		**age_low**	55	**age_high**	
age_unit	Year	**body_part_examined**		**imaging_modality**	
image_dataset_type		**materials**	DNA , Plasma , Serum	**storage_temperatures**	
sample_access_fee		**sample_access_joint_project**		**sample_access_description**	Blood is available for new determinations for specific

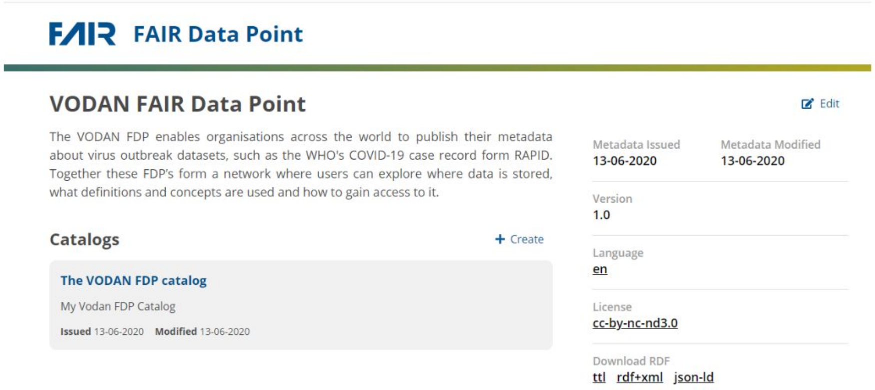
structured metadata following a scheme ... but how machine accessible are they?

FAIR Data Point

Software to create and expose metadata for datasets.

Developed in Java by GO-FAIR, and later in python by NL eScience Center.

Docker deployable; has standardized metadata, plans to extend to arbitrary metadata schemes



FAIR Data Point

VODAN FAIR Data Point

Edit

The VODAN FDP enables organisations across the world to publish their metadata about virus outbreak datasets, such as the WHO's COVID-19 case record form RAPID. Together these FDP's form a network where users can explore where data is stored, what definitions and concepts are used and how to gain access to it.

Metadata Issued	Metadata Modified
13-06-2020	13-06-2020

Version
1.0

Language
en

License
cc-by-nc-nd3.0

Download RDF
ttl rdf+xml json-ld

Catalogs

+ Create

The VODAN FDP catalog

My Vodan FDP Catalog

Issued 13-06-2020 **Modified** 13-06-2020

Maastricht University

Institute of Data Science

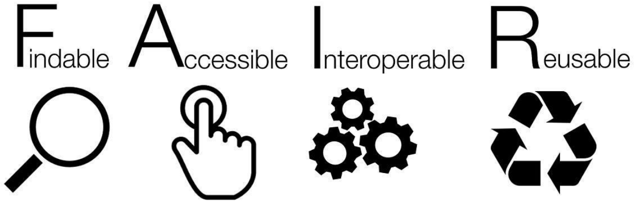
https://www.fairdatapoint.org/

Medical Informatics Europe MIE2022 EFMI

Findable Accessible Interoperable Reusable

However, the FAIR Principles indicate the functional requirements, but do not specify the technical implementation details

How do I know if my data is FAIR or FAIR enough?



Findable Accessible Interoperable Reusable

However, the FAIR Principles indicate the functional requirements, but do not specify the technical implementation details

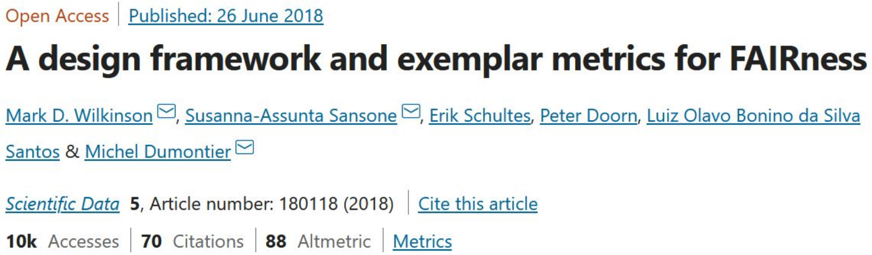
↓

Need to evaluate FAIRness, where there is a number of possible implementations...

Measuring FAIRness

A framework for defining evaluative metrics. Every metric should be coupled with a document that describes **what is being measured**, **why** one wants to measure it, **what a valid result is** and **how** one obtains it.

The metric descriptions should be **clear**, **realistic**, **discriminating**, **measureable**, and be **universally** applicable.



Open Access | Published: 26 June 2018

A design framework and exemplar metrics for FAIRness

Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos & Michel Dumontier

Scientific Data **5**, Article number: 180118 (2018) | Cite this article

10k Accesses | **70** Citations | **88** Altmetric | Metrics

14 universal metrics covering each of the FAIR sub-principles. The **metrics demand evidence** from the community, some of which may require specific new action

Maastricht University

Institute of Data Science

Medical Informatics Europe

MIE2022

EFMI

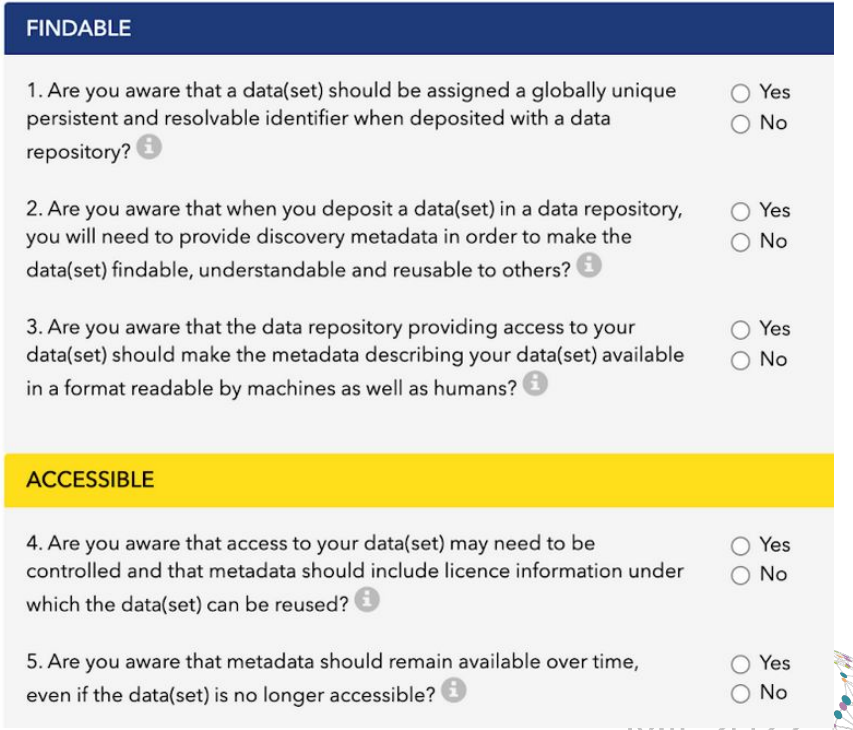
Assessment Modalities

- Manual evaluation
- Semi-automated evaluation
- Automated evaluation

FAIRassist: *https://fairassist.org*

Assessment Modalities

- **Manual evaluation**
  - Extensive flexibility to explore both quantitative and qualitative aspects
  - Takes a lot of time to perform the evaluation
  - Can a human really evaluate if a resource is machine-readable?
  - FAIRdat, FAIR-aware, **DMP**
- **Semi-automated evaluation**
- **Automated evaluation**



FINDABLE

1. Are you aware that a data(set) should be assigned a globally unique persistent and resolvable identifier when deposited with a data repository? ⓘ
○ Yes
○ No

2. Are you aware that when you deposit a data(set) in a data repository, you will need to provide discovery metadata in order to make the data(set) findable, understandable and reusable to others? ⓘ
○ Yes
○ No

3. Are you aware that the data repository providing access to your data(set) should make the metadata describing your data(set) available in a format readable by machines as well as humans? ⓘ
○ Yes
○ No

ACCESSIBLE

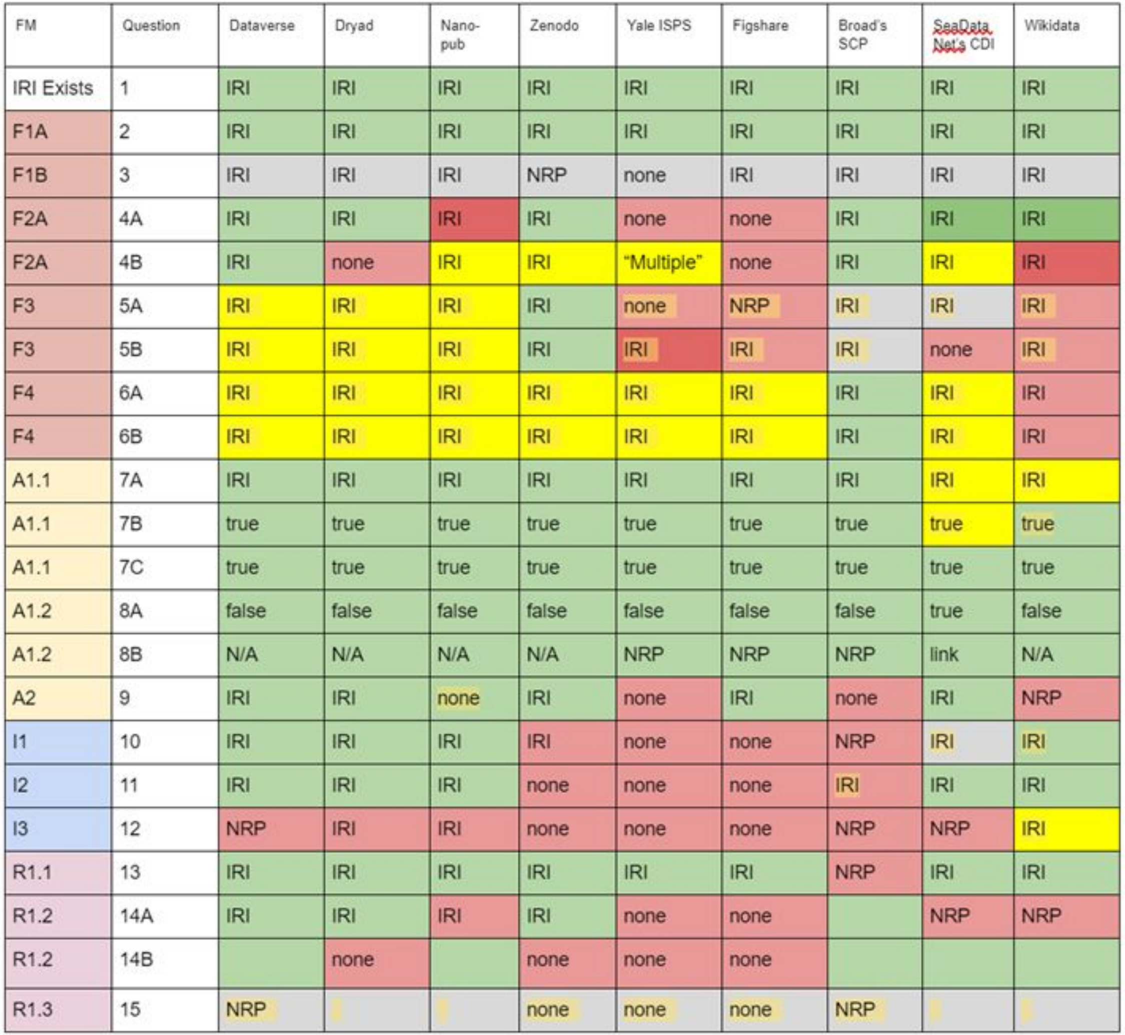
4. Are you aware that access to your data(set) may need to be controlled and that metadata should include licence information under which the data(set) can be reused? ⓘ
○ Yes
○ No

5. Are you aware that metadata should remain available over time, even if the data(set) is no longer accessible? ⓘ
○ Yes
○ No

Table 2. Summary of FAIR metrics self-scoring.

Green = passes FAIR Metric
Red = fails FAIR Metric
Yellow = problementatic (for example, incorrectly interpreted question)
Gray = Can not be evaluated

IRI = Respondent gives an IRI
none = Respondent answered "none"
NRP = No Response Provided



FM	Question	Dataverse	Dryad	Nano-pub	Zenodo	Yale ISPS	Figshare	Broad's SCP	SeaData Net's CDI	Wikidata
IRI Exists	1	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI
F1A	2	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI
F1B	3	⬜ IRI	⬜ IRI	⬜ IRI	⬜ NRP	⬜ none	⬜ IRI	⬜ IRI	⬜ IRI	⬜ IRI
F2A	4A	🟩 IRI	🟩 IRI	🟥 IRI	🟩 IRI	🟥 none	🟥 none	🟩 IRI	🟩 IRI	🟩 IRI
F2A	4B	🟩 IRI	🟥 none	🟨 IRI	🟨 IRI	🟨 "Multiple"	🟥 none	🟩 IRI	🟨 IRI	🟥 IRI
F3	5A	🟨 IRI	🟨 IRI	🟨 IRI	🟩 IRI	🟥 none	🟥 NRP	⬜ IRI	⬜ IRI	🟥 IRI
F3	5B	🟨 IRI	🟨 IRI	🟨 IRI	🟩 IRI	🟥 IRI	🟥 IRI	⬜ IRI	⬜ none	🟥 IRI
F4	6A	🟨 IRI	🟨 IRI	🟨 IRI	🟨 IRI	🟨 IRI	🟨 IRI	🟩 IRI	🟨 IRI	🟥 IRI
F4	6B	🟨 IRI	🟨 IRI	🟨 IRI	🟨 IRI	🟨 IRI	🟨 IRI	🟩 IRI	🟨 IRI	🟥 IRI
A1.1	7A	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟨 IRI	🟨 IRI
A1.1	7B	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟨 true	🟩 true
A1.1	7C	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true	🟩 true
A1.2	8A	🟩 false	🟩 false	🟩 false	🟩 false	🟩 false	🟩 false	🟩 false	🟩 true	🟩 false
A1.2	8B	🟩 N/A	🟩 N/A	🟩 N/A	🟩 N/A	🟩 NRP	🟩 NRP	🟩 NRP	🟩 link	🟩 N/A
A2	9	🟩 IRI	🟩 IRI	🟩 none	🟩 IRI	🟥 none	🟩 IRI	🟥 none	🟩 IRI	🟥 NRP
I1	10	🟩 IRI	🟩 IRI	🟩 IRI	🟥 IRI	🟥 none	🟥 none	🟥 NRP	⬜ IRI	🟩 IRI
I2	11	🟩 IRI	🟩 IRI	🟩 IRI	🟥 none	🟥 none	🟥 none	🟥 IRI	🟩 IRI	🟩 IRI
I3	12	🟥 NRP	🟥 IRI	🟥 IRI	🟥 none	🟥 none	🟥 none	🟥 NRP	🟥 NRP	🟨 IRI
R1.1	13	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟩 IRI	🟥 NRP	🟩 IRI	🟩 IRI
R1.2	14A	🟩 IRI	🟩 IRI	🟥 IRI	🟩 IRI	🟥 none	🟥 none	🟩	🟥 NRP	🟥 NRP
R1.2	14B	🟩	🟥 none	🟩	🟥 none	🟥 none	🟥 none	🟩	🟩	🟩
R1.3	15	⬜ NRP	⬜	⬜	⬜ none	⬜ none	⬜ none	⬜ NRP	⬜	⬜

Findings

- Promising first assessments
- Conflicting reporting in Findability
- Biggest issues around interoperability and provenance

Maastricht University
Institute of Data Science

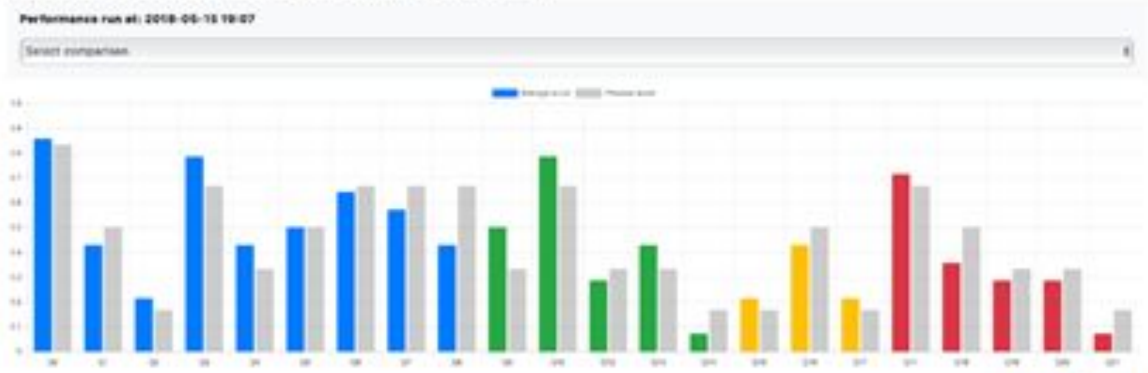
Bio-IT World
CONFERENCE & EXPO '18

https://purplepolarbear.nl

56 users
32 datasets
25 metric evaluations

PPB FAIR Score	Before	After	Δ
cBioPortal	39	59	+20
JaxTeam	26	54	+28
Broad Single Cell	29	52	+23
Bioassay	48	75	+27

Metric evaluator: Manual check


Performance run at: 2018-05-15 19:07

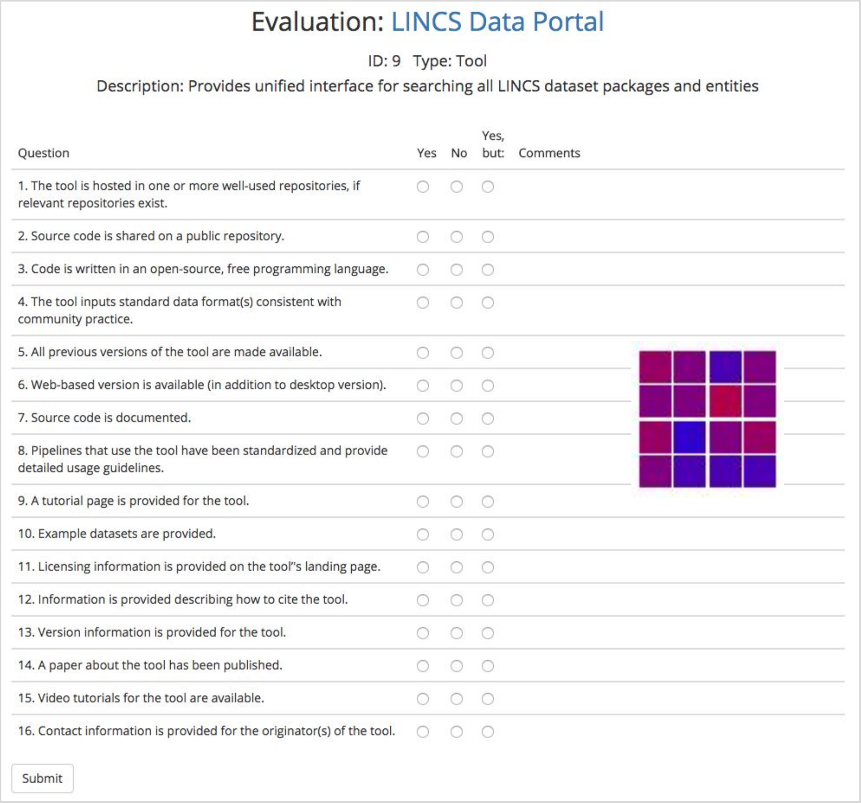
Select comparison

GO FAIR

@micheldumontier::FAIRCDR:2018-10-01

Assessment Modalities

- Manual evaluation
- **Semi-automated evaluation**
  - Combine objective and subjective assessments
  - Automatically retrieve standardized metadata for online resources (e.g. FAIRSharing)
  - FAIRShake
- Automated evaluation



Evaluation: LINCS Data Portal

ID: 9 Type: Tool

Description: Provides unified interface for searching all LINCS dataset packages and entities

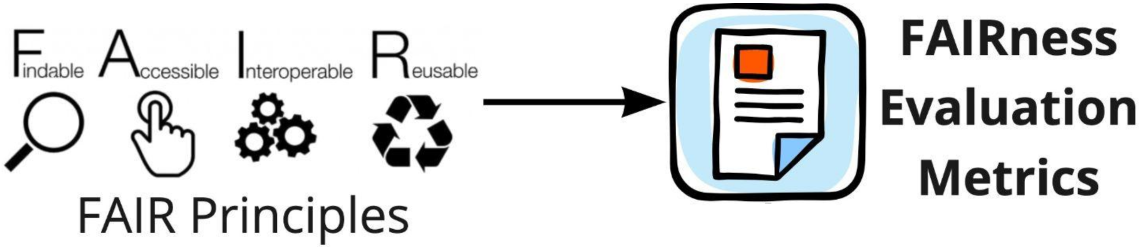
Question	Yes	No	Yes, but:	Comments
1. The tool is hosted in one or more well-used repositories, if relevant repositories exist.	○	○	○	
2. Source code is shared on a public repository.	○	○	○	
3. Code is written in an open-source, free programming language.	○	○	○	
4. The tool inputs standard data format(s) consistent with community practice.	○	○	○	
5. All previous versions of the tool are made available.	○	○	○	
6. Web-based version is available (in addition to desktop version).	○	○	○	
7. Source code is documented.	○	○	○	
8. Pipelines that use the tool have been standardized and provide detailed usage guidelines.	○	○	○	
9. A tutorial page is provided for the tool.	○	○	○	
10. Example datasets are provided.	○	○	○	
11. Licensing information is provided on the tool"s landing page.	○	○	○	
12. Information is provided describing how to cite the tool.	○	○	○	
13. Version information is provided for the tool.	○	○	○	
14. A paper about the tool has been published.	○	○	○	
15. Video tutorials for the tool are available.	○	○	○	
16. Contact information is provided for the originator(s) of the tool.	○	○	○	

Assessment Modalities

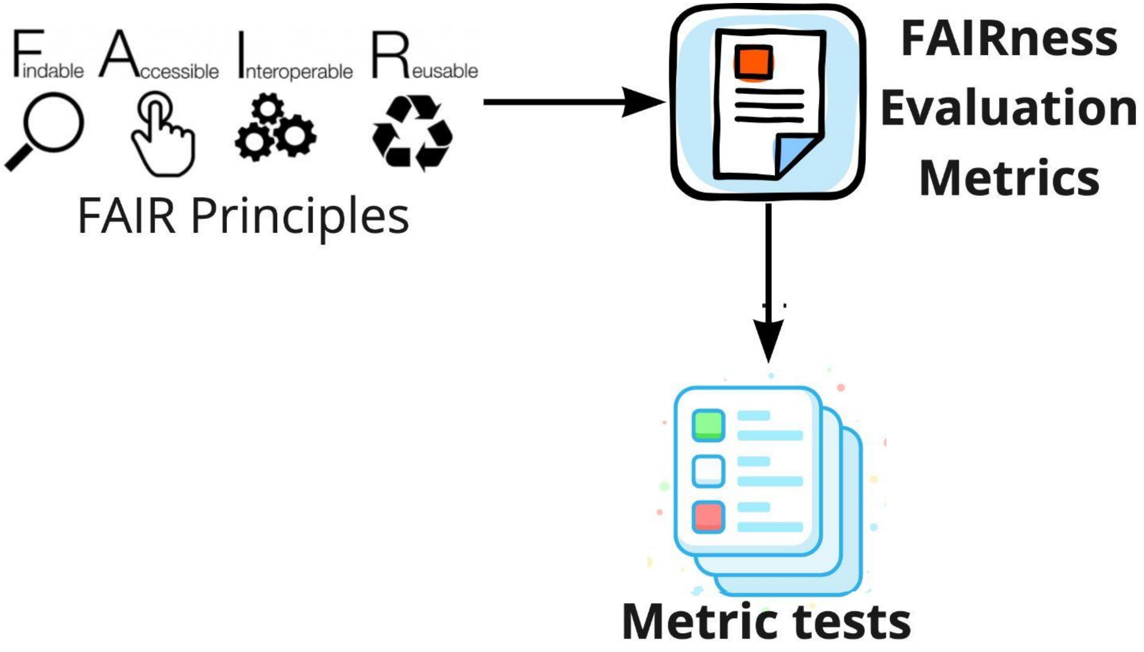
- Manual evaluation
- Semi-automated evaluation
- **Automated evaluation**
  - Efficient approach for evaluation
  - Requires all relevant information to be available to a machine
  - Flexibility in selection and application of metrics
  - Implemented as a web application that takes in persistent identifier and produces a report

Automated FAIR Evaluation workflow

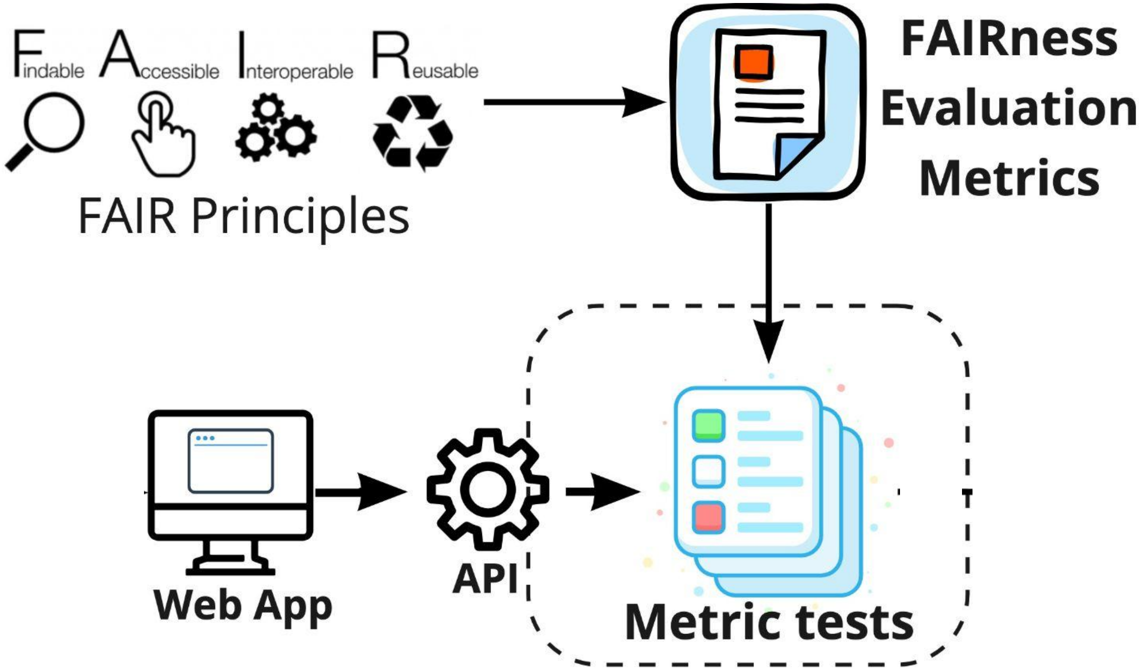
Automated FAIR Evaluation workflow



Automated FAIR Evaluation workflow



Automated FAIR Evaluation workflow

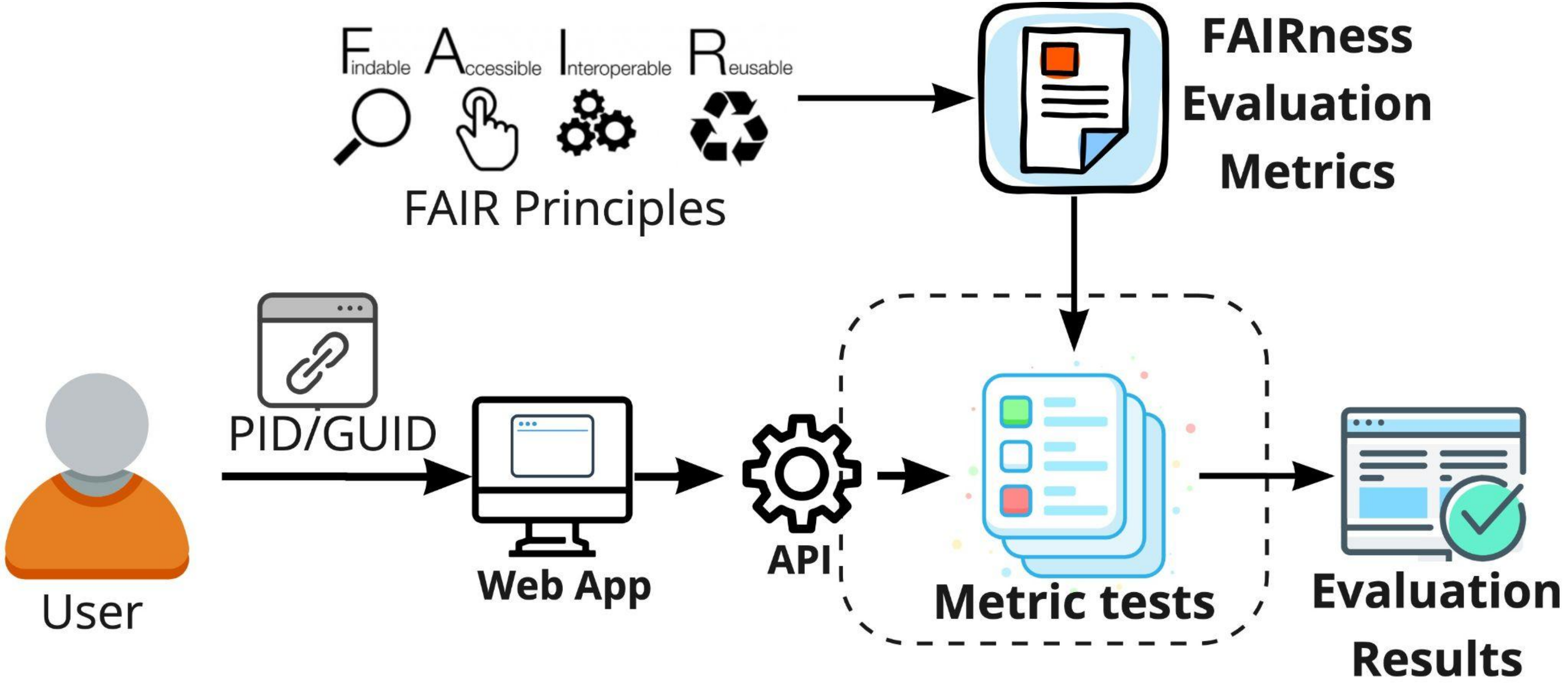

Findable Accessible Interoperable Reusable
FAIR Principles
FAIRness Evaluation Metrics
Web App
API
Metric tests

Maastricht University

Institute of Data Science

Medical Informatics Europe MIE2022 EFMI

Automated FAIR Evaluation workflow

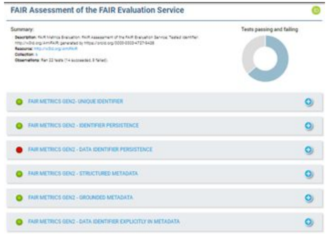

Findable Accessible Interoperable Reusable
FAIR Principles
FAIRness Evaluation Metrics
PID/GUID
User
Web App
API
Metric tests
Evaluation Results

Maastricht University

Institute of Data Science

Medical Informatics Europe MIE2022 EFMI

Automated FAIR Evaluation Tools



FAIR Assessment of the FAIR Evaluation Service

Summary:

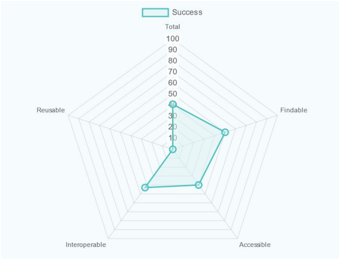
Description: FAIR Metrics Evaluation: FAIR Assessment of the FAIR Evaluation Service; Tested identifier http://w3id.org/AmIFAIR; generated by https://orcid.org/0000-0003-4727-9428
Resource: http://w3id.org/AmIFAIR
Collection: s
Observations: Ran 22 tests (14 succeeded, 8 failed)

Tests passing and failing

- ● FAIR METRICS GEN2- UNIQUE IDENTIFIER
- ● FAIR METRICS GEN2 - IDENTIFIER PERSISTENCE
- ● FAIR METRICS GEN2 - DATA IDENTIFIER PERSISTENCE
- ● FAIR METRICS GEN2 - STRUCTURED METADATA
- ● FAIR METRICS GEN2 - GROUNDED METADATA
- ● FAIR METRICS GEN2 - DATA IDENTIFIER EXPLICITLY IN METADATA

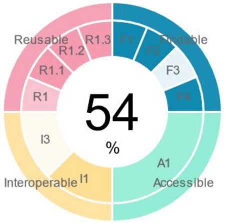
FAIR Evaluator

https://w3id.org/AmIFAIR



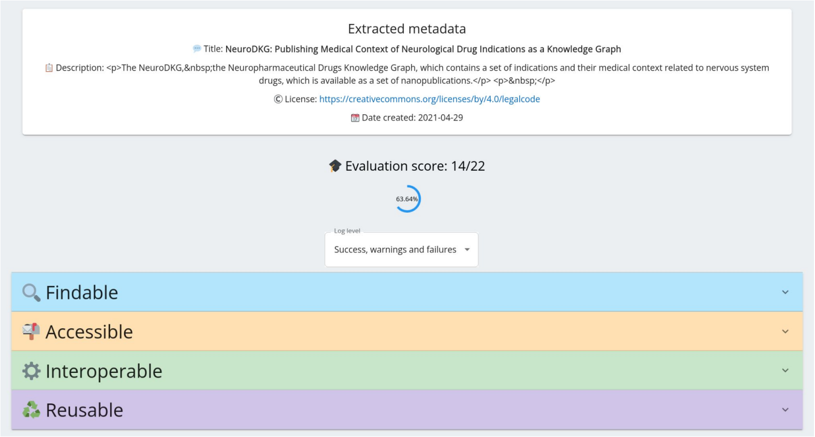
FAIR Checker

https://fair-checker.france-bioinformatique.fr/base_metrics



F-UJI

https://www.f-uji.net



Extracted metadata

💬 Title: NeuroDKG: Publishing Medical Context of Neurological Drug Indications as a Knowledge Graph

📋 Description: <p>The NeuroDKG, the Neuropharmaceutical Drugs Knowledge Graph, which contains a set of indications and their medical context related to nervous system drugs, which is available as a set of nanopublications.</p> <p>&nbsp;</p>

© License: https://creativecommons.org/licenses/by/4.0/legalcode

📅 Date created: 2021-04-29

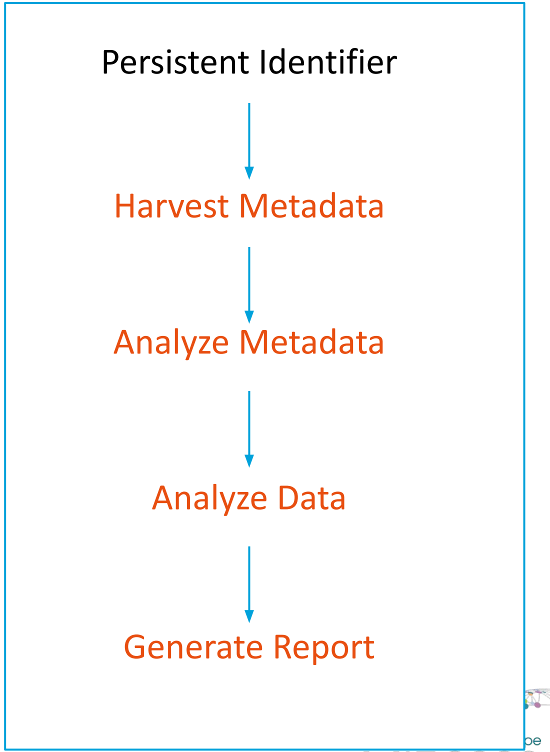
🎓 Evaluation score: 14/22

FAIR Enough

https://w3id.org/fair-enough

The evaluation tools may generate different FAIRness assessment results

- on characteristics of the **evaluation tools**
  - harvest different metadata
- on the FAIRness evaluation **metrics**
  - different way to analyze metadata and data
- on the evaluation **results**
  - using different scoring system, generate different results



Persistent Identifier

Harvest Metadata

Analyze Metadata

Analyze Data

Generate Report

Metadata Harvesting process

1. Try to extract metadata from the HTML page
2. Use HTTP requests with content-negotiation to ask for the data in a specific format (RDF, JSON-LD)
3. Check for “Signposting” links redirection (aka. Web Linking) in the response headers, follow the redirection and repeat the previous steps

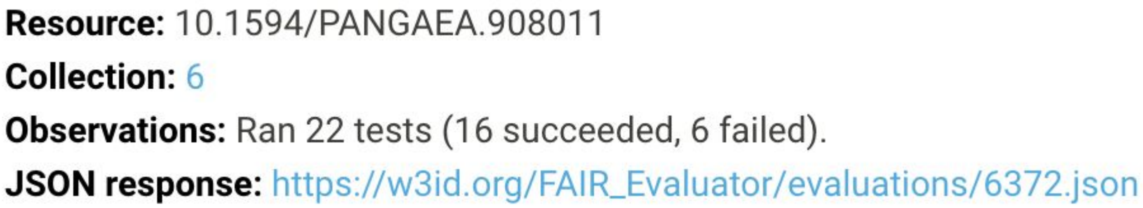
Variations in using different FAIR evaluators

1. Choice of identifier matters : DOI vs URL

DOI vs URL

DOI: 10.1594/PANGAEA.908011

Description: Metric to test if the metadata contains the unique identifier to the metadata itself



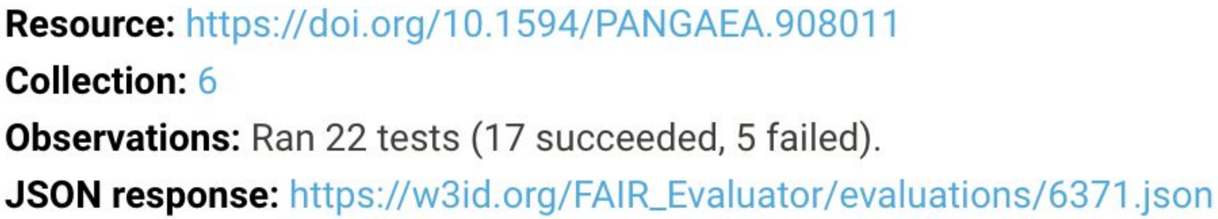
Resource: 10.1594/PANGAEA.908011

Collection: 6

Observations: Ran 22 tests (16 succeeded, 6 failed).

JSON response: https://w3id.org/FAIR_Evaluator/evaluations/6372.json

URL: https://doi.org/10.1594/PANGAEA.908011



Resource: https://doi.org/10.1594/PANGAEA.908011

Collection: 6

Observations: Ran 22 tests (17 succeeded, 5 failed).

JSON response: https://w3id.org/FAIR_Evaluator/evaluations/6371.json

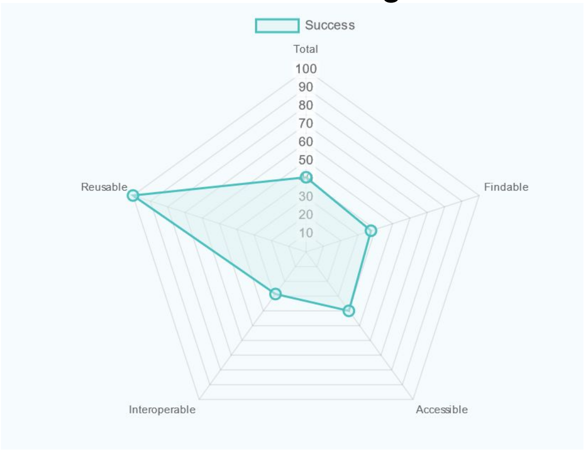
Variations in using different FAIR evaluators

1. Choice of identifier matters : DOI vs URL

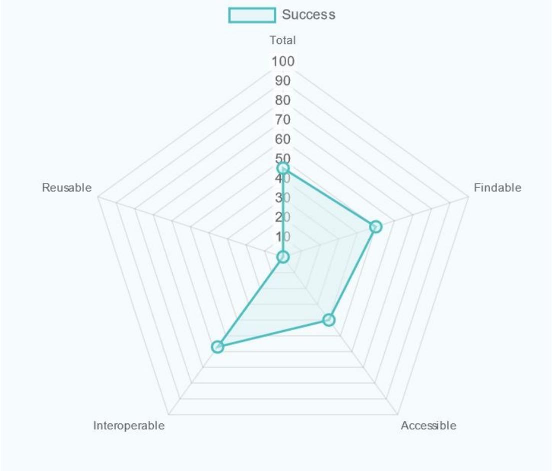
2. Most repositories don’t present structured metadata about themselves

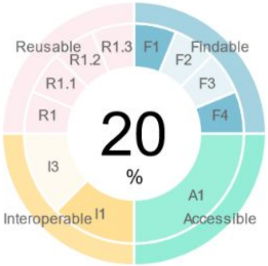
Wikidata.org

DOI:10.25504/FAIRsharing.6s749p



FAIR checker





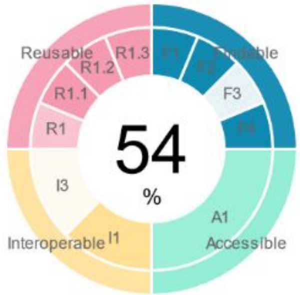
Findable: 2 of 7

Accessible: 1 of 3

Interoperable: 2 of 4

Reusable: 0 of 10

F-UJI



Findable: 6 of 7

Accessible: 1 of 3

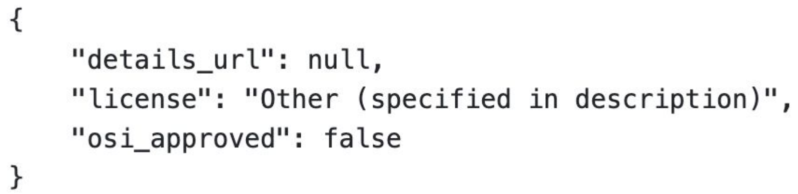
Interoperable: 2 of 4

Reusable: 4 of 10

Variations in using different FAIR evaluators

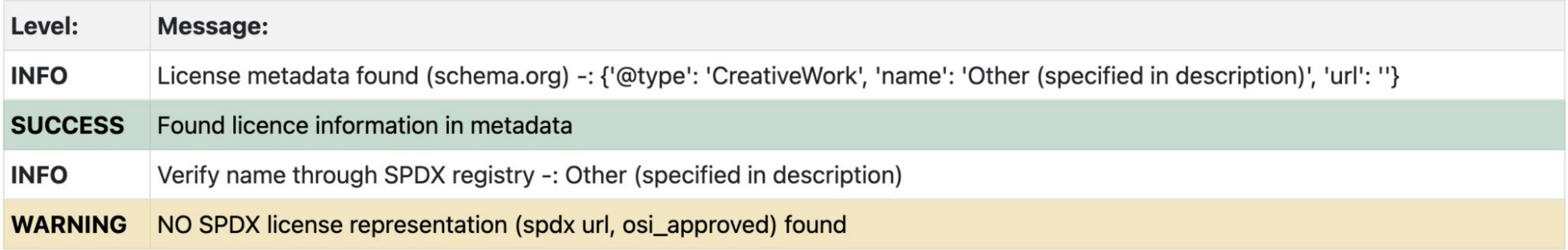
1. Choice of identifier matters : DOI vs URL
2. Most repositories don’t present structured metadata about themselves
3. Some tools do not check the metadata (license) closely

Test on COVID-19 Open Research Dataset Challenge (CORD-19)



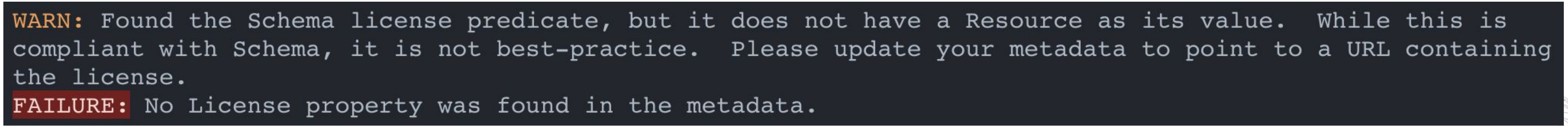
```
{
    "details_url": null,
    "license": "Other (specified in description)",
    "osi_approved": false
}
```

F-UJI



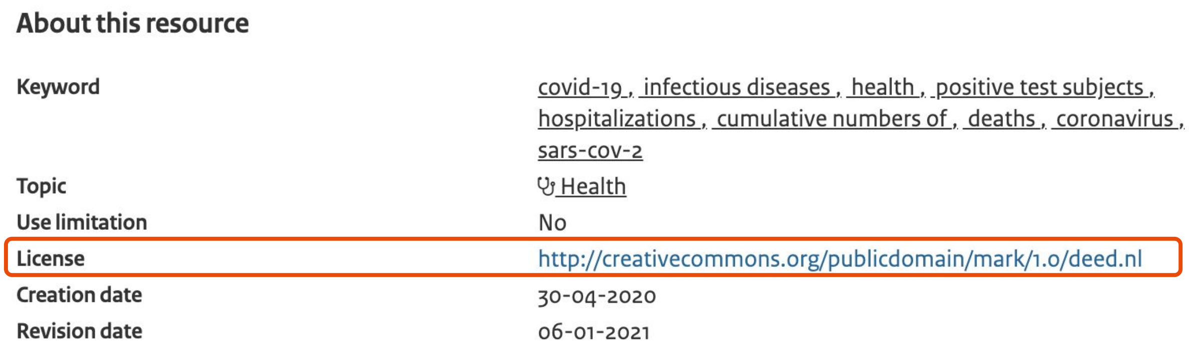
Level:	Message:
INFO	License metadata found (schema.org) -: {'@type': 'CreativeWork', 'name': 'Other (specified in description)', 'url': ''}
SUCCESS	Found licence information in metadata
INFO	Verify name through SPDX registry -: Other (specified in description)
WARNING	NO SPDX license representation (spdx url, osi_approved) found

FAIR Evaluator



```
WARN: Found the Schema license predicate, but it does not have a Resource as its value.  While this is
compliant with Schema, it is not best-practice.  Please update your metadata to point to a URL containing
the license.
FAILURE: No License property was found in the metadata.
```

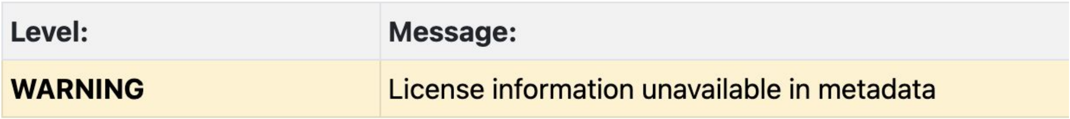
Dutch COVID cases dataset (National Institute for Public Health and the Environment)



About this resource

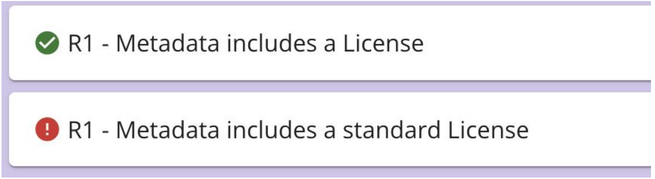
Keyword	covid-19 , infectious diseases , health , positive test subjects , hospitalizations , cumulative numbers of , deaths , coronavirus , sars-cov-2
Topic	Health
Use limitation	No
License	http://creativecommons.org/publicdomain/mark/1.0/deed.nl
Creation date	30-04-2020
Revision date	06-01-2021

F-UJI



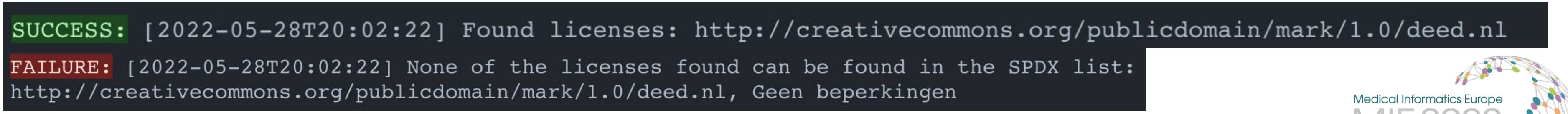
Level:	Message:
WARNING	License information unavailable in metadata

FAIR Evaluator



✓ R1 - Metadata includes a License

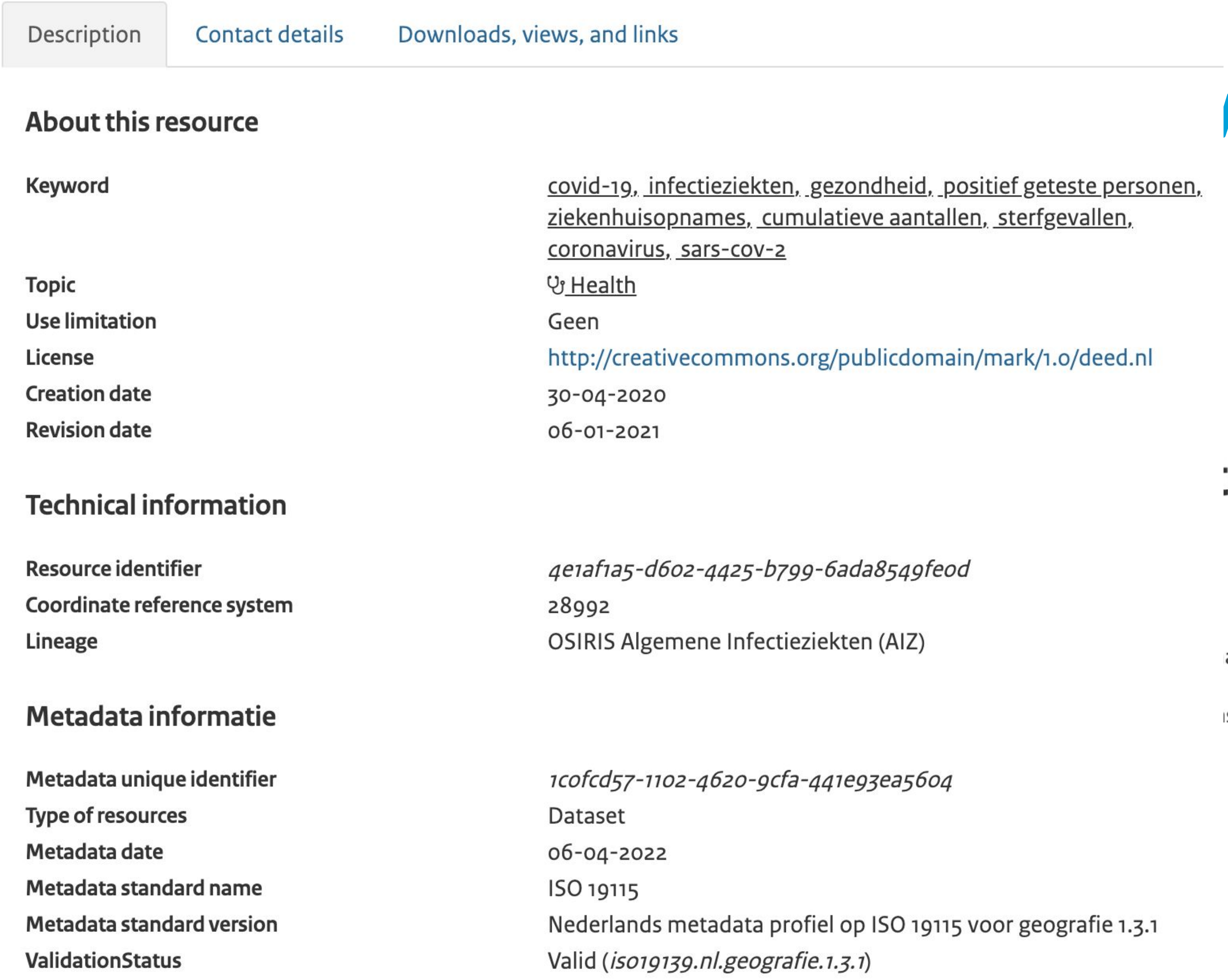
! R1 - Metadata includes a standard License



```
SUCCESS: [2022-05-28T20:02:22] Found licenses: http://creativecommons.org/publicdomain/mark/1.0/deed.nl
FAILURE: [2022-05-28T20:02:22] None of the licenses found can be found in the SPDX list:
http://creativecommons.org/publicdomain/mark/1.0/deed.nl, Geen beperkingen
```

FAIR evaluations are subjective

Some FAIR requirements are generic, but what makes a resource FAIR depend on the domain, type of resource evaluated, and community in which this resource is used



About this resource

Keyword	covid-19, infectieziekten, gezondheid, positief geteste personen, ziekenhuisopnames, cumulatieve aantallen, sterfgevallen, coronavirus, sars-cov-2
Topic	Health
Use limitation	Geen
License	http://creativecommons.org/publicdomain/mark/1.0/deed.nl
Creation date	30-04-2020
Revision date	06-01-2021

Technical information

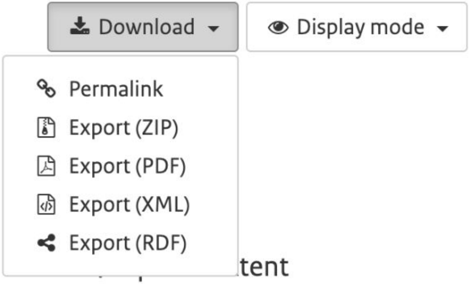
Resource identifier	*4e1af1a5-d602-4425-b799-6ada8549feod*
Coordinate reference system	28992
Lineage	OSIRIS Algemene Infectieziekten (AIZ)

Metadata informatie

Metadata unique identifier	*1cofcd57-1102-4620-9cfa-441e93ea5604*
Type of resources	Dataset
Metadata date	06-04-2022
Metadata standard name	ISO 19115
Metadata standard version	Nederlands metadata profiel op ISO 19115 voor geografie 1.3.1
ValidationStatus	Valid (*iso19139.nl.geografie.1.3.1*)

ve

te



tent

af

s ziin

Spatial extent

Summary

- FAIR is really about providing structured data and metadata in a manner that machines can find and decipher
- Making FAIR data is dataset specific - there are several community-driven guides to specify the details e.g. identifiers, standards, repositories, licenses, etc
- Several FAIR Evaluator tools exist, but these vary in their performance until such time that *they conform to a standard.*

Part 2 - Assess FAIRness of select biomedical resources

Example evaluation

Evaluate a dataset about Cell lines:
https://w3id.org/ejp-rd/fairdatapoints/wp13/dataset/c5414323-eab1-483f-a883-77951f246972

(Short URL: https://bit.ly/miefairdata)

Using the **FAIR Maturity Indicator for Rare Disease** collection, to see if this dataset conforms to all requirements of a specific community doing research on rare diseases

Rare Disease FAIR maturity indicators

Simple collection doing 2 tests for a specific community:

1. Validate the resource metadata is machine readable, and complies with a specific schema
2. Check if the resource metadata can be found in a specific search engine (the FAIR Data Point index in this case)

Evaluation of BBMRI resource

https://catalogue.bbmri.nl/menu/main/dataexplorer/details/eu_bbmri_eric_collections/bbmri-eric:ID:NL_AAAACXPRCP2M6ACQK2ME25QAAE:collection:35

(Short URL: https://bit.ly/miefairdata2)

Use the **fair-evaluator-maturity-indicators** collection, a more complex collection doing 22 generic tests

Evaluate your resource

1. Go to https://fair-enough.semanticscience.org **(Short URL: https://bit.ly/fairenoughtool)**
2. Select the collection of metrics tests you want to use to evaluate your resource
3. Paste the URL to your resource in the box saying “URL of the resource to evaluate”
4. Click “Start the evaluation”, after a few seconds you will see the results and detail of the evaluation for your resource

Define a new collection

You can also define a new collection with FAIR Metrics Tests already registered in FAIR enough:

1. Go to https://fair-enough.semanticscience.org/collection/create **(Short URL: https://bit.ly/faircollections)**
2. And login with your ORCID

Register a new Metrics test

You can also register a new Metric test that can then be added as part of a collection:

https://fair-enough.semanticscience.org/metrics

(Short URL: https://bit.ly/fairmetrics)

Part 3 - Create a custom FAIR metrics test

09:40 - 10:15

- Discuss the need and potential of community-based/domain-specific metric tests and collections, with a focus on emergent standards in the rare disease community
- Describe how custom evaluation tests can be created using fair-test library [19]
- Guide participants to create, register, and execute a domain-specific FAIRness

test and metric collection

Maturity Indicators for your community

What kind of requirements would you like to test?

Popular type of test:

- Specific metadata format
- Specific schema

A tool to make it simple

Define and deploy your FAIR tests can be easily defined and deployed using a developer friendly library: https://maastrichtu-ids.github.io/fair-test

Example of FAIR tests

For the Rare Disease community:

https://rare-disease.api.fair-enough.semanticscience.org

(Short URL: https://bit.ly/rarediseasefair)

Code:

https://github.com/LUMC-BioSemantics/RD-FAIRmetric-F4

(Short URL:https://bit.ly/faritest)

Part 4 - Discussion + Closing

10:15-10:30

- Participants share their thoughts and experience
- Future of FAIR evaluation in the biomedical informatics community

Discussion

Slides: https://bit.ly/miefaireva

- **Difficult to understand how each tool performs the evaluation without looking at source code or technical specifications.**
- **Apparent differences between the tools**
  - Different understanding of certain concepts.
  - Different depth of information extraction.
  - Different implementations of the metrics

Future Work : focus on standardized benchmarks to critically evaluate the functioning of these and future FAIRness evaluation tools.

Acknowledgements

Slides: https://bit.ly/miefaireva

Tool authors for helpful discussions:

- **FAIR Enough:** Vincent Emonet
- **FAIR Evaluator:** Mark Wilkinson , Pablo Alarcón
- **FAIR Checker:** Thomas Rosnet, Alban Gaignard
- **F-UJI** (GitHub): Robert Huber

Funding

EOSC-Life

NIH National Center for Advancing Translational Sciences

Digital Society

Maastricht University