A generic framework to better understand and compare FAIRness measures

Philippe Lamarre, Jennie Andersen, Alban Gaignard, Sylvie Cazalens

A Generic Framework to Better Understand and Compare FAIRness Measures

Philippe Lamarre[1], Jennie Andersen[1], Alban Gaignard[2], and Sylvie Cazalens[1]

[1] INSA Lyon, CNRS, Ecole Centrale de Lyon, Universite Claude Bernard Lyon 1, Université Lumière Lyon 2, LIRIS, UMR5205 69621 Villeurbanne, France
firstname.lastname@insa-lyon.fr
[2] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
alban.gaignard@univ-nantes.fr

Abstract. In recent years, the adoption of the FAIR principles has achieved notable success. This progress has led to the development of numerous assessment tools originating from diverse fields of application, thus addressing diverse object types, interpretations and implementations. Given the plethora of proposals available, it is crucial for users to precisely understand these measures, compare them effectively, make informed choices, and accurately interpret the obtained measurements. To meet these needs, we propose a model to formally represent and analyze measures. Besides the benefit of homogenization, it allows for the formal definition of three characteristic quantities: coverage, granularity and impact. Our experiments show how these quantities (i) contribute to explain different scores obtained by digital artifacts using two different state-of-the-art assessment engines, (ii) enable a comparative study of different FAIRness measures, independently of any digital artifact.

Keywords: FAIR data · FAIR assessments · Trustworthiness.

1 Introduction

These recent years, the FAIR principles [40] have been increasingly adopted to assess the Findability, Accessibility, Interoperability and Reusability of their digital resources. Due to this widespread adoption, they have been specialized or even extended to meet the needs of very different scientific communities. For instance, the RDA FAIR4RS working group derived these principles to specifically address research software [5], typically targeting their usability and reusability within other software. These principles have also been adapted in the context of AI [24], ontology development and semantic artifacts [32,9], data analysis workflows [38].

To support people in these assessment tasks, numerous tools have been developed, originating from diverse fields of application. They may address different types of objects, stem from different interpretations or implementations of the principles. Consequently, besides the varied terminology, one can notice that sometimes sub-principles are skipped, and how indicators are expressed or implemented changes from tool to tool.

In addition, when scores are provided, they may not range in the same interval; the functions used to aggregate the scores, as well as the weights assigned to the various indicators, may also differ. In some way, this diversity is understandable as the FAIR principles are not restrictive guidelines.

However, with numerous and diverse available tools, a user may be faced to different questions such as: "Why does my resource get such a score with this tool?", "Why does it receive a higher score with tool A compared to tool B?", "What are the differences between tools A and B?", "Which one fits my needs better?". To answer such questions, some studies already have proposed comparisons of tools, based on metrics used, on characteristics of the tools themselves, on the measurements obtained with numerous datasets [7,34,42,27,35,19,29]. These studies are clearly useful, but it is still a challenge to interpret scores, understand and compare FAIRness measures, and make informed choices.

This article aims to tackle this challenge based on two main points. First we observe that the FAIR principles alone do not offer a sufficient framework to take into account the multiplicity of variations found from a measure to another. A generic model enabling some homogenization in representing the measures could help. Second, to our knowledge, there are no formally defined characteristic quantities to reflect the salient features of a given measure, that could help both its understanding and its comparison with other ones.

The remainder of the paper is organized as follows. In section 3, we propose a formal model based on the tree structure of the FAIR principles, coping with the variability of FAIRness measures. Section 4 outlines the methodology for representing a given measure within this model, using FAIR-checker and F-UJI as examples. Section 5 introduces the coverage rate, granularity and impact quantities facilitating the comparison of FAIRness measures. We show with experimental results in section 6 how our framework can be used to provide better insights on diverging FAIR assessments. We propose concluding remarks in section 7.

2 Related Work

The FAIR principles were published in 2016 [40] as general guidelines for the publication of digital resources to make them Findable, Accessible, Interoperable, and Reusable. Since then, numerous methods and tools have been developed to assess FAIRness, each with its own interpretation of the principles. This variety of interpretations and of types of assessment methods (automated tools, checklists, self-assessment questionnaires, etc.) makes them difficult to compare. To address this problem, the FAIR data maturity model [4] proposes a set of indicators that express measurable aspects of the FAIR principles and on which future evaluation tools can be based. Although this initiative proposes consensual definitions adopted by large multi-disciplinary communities, it is still challenging to compare FAIRness measures in their whole, from implementations to evaluation results.

Several comparisons have already been conducted. Slamkov *et al.* [34] compare five questionnaires and checklists: ARDC's tool [3], CSIRO's tool [10], SATIFYD [11], EUDAT checklist [25], and the SHARC grid [12], according to their main characteristics (type, documentation, dependency to a specific repository, automated score computation) and the results obtained on seven datasets.

Another comparison [35] focuses on three automated tools: F-UJI [13], FAIR Evaluation Service[3] [41] and FAIR-Checker [17]. They are all compared based on distinguishing aspects (documentation, availability of the code, format and log of the results...). Then the author focus on F-UJI and the FAIR Evaluation Service to compare their metrics/indicators in detail, first by focusing on their expression in natural language and then on the experimental results obtained on three datasets. Since then, both F-UJI and FAIR-Checker have changed. Wilkinson *et al.* [42] highlight that some of the differences between the results of F-UJI and the FAIR Evaluation Service are not due to the metrics themselves, but to their different ways of collecting the metadata to be evaluated. This applies to all automated tools and contributes to the difficulty of comparison.

Krans *et al.* [27] provide a detailed qualitative comparison of ten tools, both questionnaires and automated tools. They mostly focus on their prerequisites, the ease and effort to use them, the type and quality of the outputs. They also test them on two datasets and observe a large variability in the FAIRness scores obtained, thus showing the difficulty in interpreting the questions in questionnaires and the differences in the implementations of the principles for automated tools. Candela *et al.* [7] provide an overview of twenty FAIR assessment tools, analyzing distinguishing features such as target, adaptability, methodology... They particularly document the divergence between declared intents of metrics and what is actually assessed. Some tools have also been compared in the context of domain-specific FAIRness assessment [19]. In particular, they compare the overall FAIRness score obtained on some datasets by the FAIR Evaluation Service, F-UJI, FAIRshake [8], and a self-assessment based on the FAIR data maturity model. They observe that the tools obtain scores close enough to consider that they give similar levels of FAIRness, especially if of the same kind (questionnaire or automated tool).

Recently, Moser *et al.* [29] propose a brief comparison of the FAIRness measures rather than the tools based on them. They focus on the FAIR Maturity Indicators [41], the FAIR Data Maturity Model [4], FAIRsFAIR metrics used in F-UJI [13] and FAIR metrics for EOSC [20]. They compare their numbers of indicators, and their structures: some metrics define indicators for intermediate principles (A1 and R1) while others do not. They also highlight that some of them give different importance to their indicators.

The main objective of our proposal is closer to this latter work, while we aim to push further the comparison of the scores and of the importance of each element in the measures. We propose an innovative approach with a generic model and formal definitions of characteristic quantities.

[3] With the metric collection: "All Maturity Indicator Tests as of May 8, 2019"

3 A Generic Model to Represent FAIRness Measures

Our aim is to define a simple unifying framework expressive enough to represent and compare as many measures as possible. We focus on the importance they give to the FAIR principles and sub-principles. In this view our analysis of FAIRness measures has identified three notions to describe their tree-like organization: principles, indicators and implementations. As the ways scores are computed vary a lot, we propose a generic representation. It ensures that if a score is computed for a given principle or sub-principle by a measure, the score computed through its representation in the model is the same. Hence, we propose to represent a measure of FAIRness as a tuple

$$\mathcal{M} = (V, E, \text{FAIR}, \Diamond, w, v_{max}, D)$$

where elements (V, E, FAIR) refer to the structure and $(\Diamond, w, v_{max}, D)$ to the score computation. They are detailed in the following.

3.1 Modeling of the Structure

Obviously, the different measures of FAIRness rely on the hierarchy of the FAIR principles [40]. Here, the term *principle* is used in a broad sense, i.e. it refers to both principles and sub-principles. We represent them as a tree, illustrated with the tree of ellipses in Figure 1. Its sets of nodes is denoted $P = \{\text{FAIR}, \text{F}, \text{F1}, \text{F2}, \text{F3}, \text{F4}, \text{A}, \text{A1}, \text{A1.1} \ldots\}$ and its edges, $E_P = \{(\text{FAIR}, \text{F}), (\text{F}, \text{F1}), (\text{F}, \text{F2}) \ldots\}$.

Then, these principles are refined into several measurable criteria, expressed in natural language, which we call indicators. An indicator is named a "metric" in FAIR-Checker [17] and F-UJI [13], a "FAIRness assessment question" in O'FAIRe [1], a "maturity indicator test" in the FAIR Evaluation Service [41], or a "check" in FOOPS! [18]. Given $\mathcal{M}$, a measure of FAIRness, we denote $I(\mathcal{M})$ its set of indicators. Finally, in the case of an automated tool, indicators are associated to implementations, belonging to set $Imp(\mathcal{M})$, allowing a resource to be evaluated on them.

Hence, the structure of a measure $\mathcal{M}$ is represented by a directed rooted tree (V, E, FAIR), simply adding indicators and implementations to the FAIR principle tree, where:

- FAIR is the root of $\mathcal{M}$ and of the FAIR principles tree;
- V are the nodes of the tree, such that $V = P \cup I(\mathcal{M}) \cup Imp(\mathcal{M})$;
- E are the edges, where $E \subseteq E_P \cup (P \times I(\mathcal{M})) \cup (I(\mathcal{M}) \times Imp(\mathcal{M}))$;
- there can only be one implementation per indicator.

In this model, an implementation can be linked only to an indicator, which in turn can be linked only to a principle, any principle, not just to the leaves of the FAIR principles tree. This is intended to simplify the representation and to ease understanding and comparison of the measures. Methods for representing existing measures that do not comply with these constraints (e.g. with hierarchies of indicators, or additional sub-principles.) are discussed in section 4.

To manipulate a measure such defined, we introduce the usual notions of children and of descendants: Let $n \in V$ be a node of the tree, then $\text{children}_{\mathcal{M}}(n)$ is the set of children of n in $\mathcal{M}$, and $\text{desc}_{\mathcal{M}}(n)$ is the set of descendants of n in $\mathcal{M}$.

3.2 Computing the Scores

Intuitively, we consider that, given some resource d to evaluate, the score at a node is obtained by a weighted aggregation of the scores obtained by its children. The score of an indicator comes directly from executing its implementation *imp* for d, so we assume a family of evaluation functions, eval_d such that $\text{eval}_d(imp)$ denotes the obtained score. In this view, we detail the elements $(\Diamond, w, v_{max}, D)$ of the representation of a measure.

- $\Diamond : \mathcal{P}(\mathbb{R}^+ \times \mathbb{R}^+) \to \mathbb{R}^+$ is an aggregation function, producing a new (aggregated) score from some pairs (*score*, *weight*). It can be either a weighted sum (noted SUM) or a weighted average (noted AVG). It returns 0 in case of an empty set.
- $w : V \to \mathbb{R}^+$ is a weighting function. Given a node n, $w(n)$ represents the importance of n with respect to its siblings in the hierarchy.
- $v_{max} : Imp(\mathcal{M}) \cup I \to \mathbb{R}^+$ is a function where, with i being an implementation or and indicator, $v_{max}(i)$ is the maximum value that can be obtained for i whatever the resource. For example, in F-UJI, $v_{max}(\text{R1-01MD}) = 4$.
- D is a function such that, given any implementation *imp*, $D(imp)$ is a set of expressions of the form: $\forall d, \text{eval}_d(imp) \geqslant v \Rightarrow \text{eval}_d(imp') \geqslant v'$. Intuitively, for any resource, if executing *imp* results in a value above v, one is warranted that implementation imp' results in a value above some minimal value v'. This is what we call a dependency between *imp* and imp'.

Keeping these notations, the computation of the score obtained by some resource d can then be expressed as follows.

$$\text{score}(\mathcal{M}, d) = \text{score}(\mathcal{M}, \text{eval}_d, \text{FAIR}) \tag{1}$$

Function score, for a given node $n \in V$, is expressed for any function eval by:

$$\text{score}(\mathcal{M}, \text{eval}, n) = \begin{cases} \text{eval}(n) & \text{if } n \in Imp(\mathcal{M}) \\ \underset{n' \in \text{children}_{\mathcal{M}}(n)}{\Diamond} (w(n'), \text{score}(\mathcal{M}, \text{eval}, n')) & \text{otherwise} \end{cases} \tag{2}$$

4 Strategies for Representing Measures in our Approach

Some measures fit the proposed model directly. For others, several strategies may be applied. We explain and illustrate them with several FAIRness measures or tools, five automated tools and two questionnaires. Two tools can be used to

assess any digital resources: FAIR-Checker [15,17], and FAIR Evaluation Service [2,41] (the indicators from the collection "All Maturity Indicator Tests as of May 8, 2019"). One tool focuses on datasets: F-UJI [14,13]. Two tools are designed specifically for ontologies: FOOPS! [16,18] and O'FAIRe [30,1]. We also consider two online questionnaires with ARDC's questionnaire [3] and SATIFYD [11], allowing to make a self-assessment of a digital resource.

All these tools can be used online, and, except for O'FAIRe, they allow to assess any digital resource. We rely preferably on the online tools of the measures since they are more up-to-date and since it is sometimes easier to understand how the global score is computed. We first discuss the representation of the structure, then the computation of the scores.

4.1 Representing the Structure of Measures

The backbone of the structure chosen to represent measures is the full FAIR principles tree. In addition, several indicators can be attached to a same principle, an indicator can only be attached to a single principle, and an implementation to a single indicator.

For measures that fully or partially use the FAIR principles tree without additional principles, the indicators remain linked to principles just as they are. Notice that the entire FAIR principles tree is kept in the representation of the measure. This applies for example to particular questionnaires which are not detailed according to each sub-principle, but only according to F, A, I and R. Some measures define and use new sub-principles with associated indicators. These sub-principles are not represented but their indicators are considered and linked to their closest parent in the FAIR principles tree.

Some measures have several level of refinement for their indicators. For instance, Wilkinson *et al.* [41] propose a level of "maturity indicator", and then of "maturity indicator test", F-UJI [13] has a level of "metrics" and then of "tests".

To map these measures to our model, we reduce them to a single level, focusing on the one highlighted by their measuring tools, the one for which a score is clearly given. Hence, in F-UJI, we keep their "metrics" as indicators, and in [41] the "maturity indicator test" since it is the only one to appear clearly in their online tool.

Notice that, even though some of the particularities of some measures are not reproduced in their representation within the generic model, the strategies maintain the characteristic elements necessary for understanding the importance they give to the FAIR principles. This is also the case for their score calculation which we explain in the following.

4.2 Representing the Scores Computations

In our model, the scoring method is a weighted aggregation of the evaluations and scores obtained all along the tree structure, as explained in section 3.2. We seek to represent the scoring of existing measures in this way, ensuring that if a

score is computed for a given (sub-)principle by the measure, the score computed through its representation in the generic model is the same.

Measures that compute a score already have maximum values $v_{max}(i)$ for implementations and indicators, an aggregation function $\Diamond$, and possibly a set D of dependencies between implementations. These elements are all kept as is in the representation of the measure. Only the weights remain to be defined.

We first explain how to determine the weight of each implementation and indicator, and then how to deduce the weights on the other nodes.

First, all the implementations get a weight equal to 1. Then, the weights of the indicators may be explicit or implicit in the initial expression of a measure. When explicit, such as in O'FAIRe, where the weights correspond to their "credits", they are kept as is in the representation. When implicit, such as in FOOPS!, they are set to 1. Some approaches, for example the FAIR Evaluation Service, do not compute any score. This means that each user is supposed to choose how to use the results obtained for implementations and indicators. We assume they all have a weight equal to 1 and that the score is computed with a weighted average.

It is now possible to weight the other nodes of the representation. If the measure uses a sum to compute the overall score, the weight of other nodes is set to 1. If the measure computes the score with a (weighted) average, then the weight assigned to a node is the sum of the weights of its children. According to this, a principle to which no indicator is attached gets a weight equal to 0 (respectively 1) in case of a weighted average (respectively a sum).

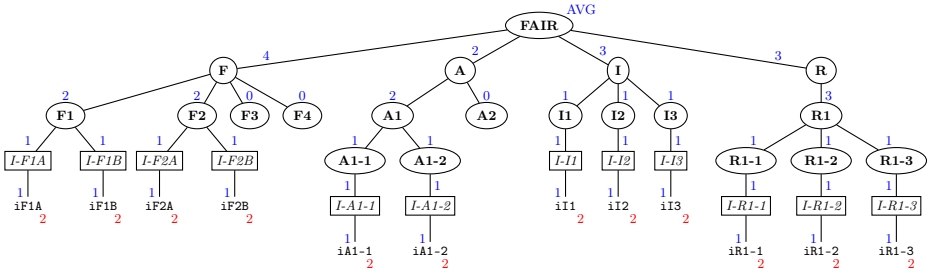
For questionnaire approaches, where the score is usually computed first on F, A, I, and R, and then globally as an average of these four scores, the weights of the main principles F, A, I, and R are 1, and the weights of the other principles are 0, since the indicators are not detailed according to them.

4.3 Examples of Representation in the Generic Model

FAIR-Checker computes the score of a resource by averaging the values obtained for the indicators. It implicitly uses weights, so we set them to 1 for all the indicators. Then, the weight of a node is assigned the sum of the weights of its children. This is illustrated in Figure 1a: an example of dependency in FAIR-checker is that the implementation of the indicator *I-I1* associated to principle I1 delegates evaluation to the one implementing *F2A* associated to F2. Hence, in Figure 1a, we represent these full delegations by equalities.

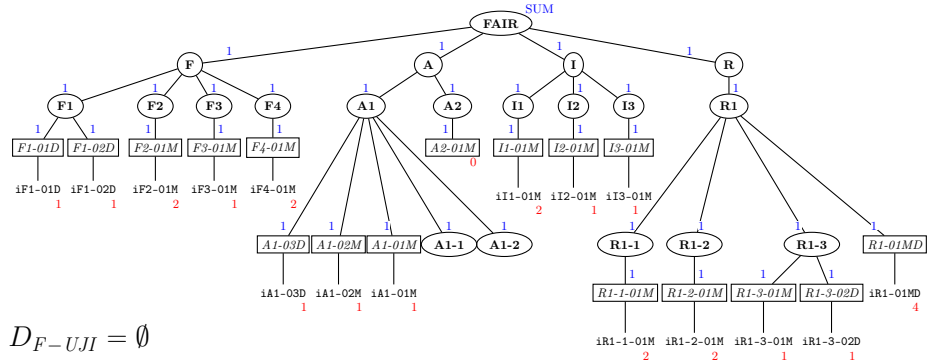
F-UJI computes the score using an unweighted sum. Its representation is illustrated in Figure 1b. Notice that in F-UJI, value v_{max} varies from 1 to 4, thus assigning different importance to indicators. Dependencies between implementations, if any, are not taken into account.

[4] According to `https://www.f-uji.net/index.php?action=methods`, assessment details of *FsF_ A2_ 01M* are excluded from the F-UJI implementation.



$D_{FAIR-Checker} = \{\forall d\, \text{eval}_d(\texttt{iI1}) = \text{eval}_d(\texttt{iF2A}), \forall d\, \text{eval}_d(\texttt{iI2}) = \text{eval}_d(\texttt{iF2B}) = \text{eval}_d(\texttt{iR1-3})\}$

(a) FAIR-Checker.



$D_{F-UJI} = \emptyset$

(b) F-UJI[4].

(p): principle – *id*: indicator – `im`: implementation

w: weight – v: v_{max} (under the leaves)

Fig. 1: Representation of measures.

5 Characteristic Quantities for Measure Analysis

To highlight the salient traits of a measure we formally define three characteristic quantities: coverage rate, granularity and impact.

Coverage Rate The role of the *coverage rate* is to measure to what extent a FAIRness measure covers the FAIR principles. Several definitions may apply. The simplest would be the proportion of leaves of the FAIR principles tree that have an indicator. However, this does not take into account the specificity of all FAIRness measures, since some of them have an indicator on A1 for instance but neither on A1.1 nor A1.2. Such measures should have a higher coverage rate than others not covering any of these three principles. Therefore, one could consider all principles and calculate the proportion of them having an indicator in their descendants. This would correct the previous problem but it may lead to

significantly high scores. For instance, with only one indicator on the principle A1.1, the coverage rate is of 20%.

Hence, we propose a coverage rate that takes into account the hierarchical aspect of the FAIR principles, where each sibling in the tree counts the same in the calculation of the coverage rate, in particular, each of the four main principles counts as 25%. Let $\mathcal{M}$ be a FAIRness measure. This *coverage rate* computed following the hierarchy is defined recursively:

$$\text{cover}(\mathcal{M}) = \text{cover}(\mathcal{M}, \text{FAIR}) \tag{3}$$

where, for a given node $n \in P$:

$$\text{cover}(\mathcal{M}, n) = \begin{cases} 0 & \text{if } \text{children}_{\mathcal{M}}(n) = \emptyset \\ \dfrac{\left(\displaystyle\sum_{n' \in \text{children}_{\mathcal{M}}(n) \cap P} \text{cover}(\mathcal{M}, n')\right) + \text{local}(\mathcal{M}, n)}{|\,\text{children}_{\mathcal{M}}(n) \cap P| + \text{local}(\mathcal{M}, n)} & \text{else} \end{cases} \tag{4}$$

with $\text{local}(\mathcal{M}, n) = 1$ if $\text{children}_{\mathcal{M}}(n) \cap I(\mathcal{M}) \neq \emptyset$ and $\text{local}(\mathcal{M}, n) = 0$ otherwise.

This definition is based on the indicators specified by the measure. Adapting it to account for implementations instead, is not particularly challenging and is not presented here. However, notice that this adaptation would produce different results in only one scenario: if at least one indicator is not implemented. Such occurrences exist and are quite natural. For instance, an indicator may prove to be unimplementable.

Granularity *Granularity* complements the previous definition. Intuitively, granularity evaluates the extent to which the indicators provide a detailed description of each principle. Higher granularity indicates a more thorough exploration of the principles and a finer-grained analysis of the FAIRness of the resources. Practically, we quantify granularity as the average number of indicators per principle that has at least one indicator.

$$\text{gran}(\mathcal{M}) = \text{gran}(\mathcal{M}, \text{FAIR}) \tag{5}$$

where, for $n \in P$:

$$\text{gran}(\mathcal{M}, n) = \frac{|\,\text{desc}_{\mathcal{M}}(n) \cap I(\mathcal{M})|}{|\{p \in (\{n\} \cup \text{desc}_{\mathcal{M}}(n)) \cap P, \text{children}_{\mathcal{M}}(p) \cap I(\mathcal{M}) \neq \emptyset\}|} \tag{6}$$

Similarly to the coverage rate, granularity is based on the indicators specified by the measure and can be easily adapted to account for implementations. The results provided by these two definitions diverge in the same scenarios as those presented for coverage rate.

Impact Last but not least, the *impact* of a principle intuitively quantifies the percentage of score that a digital element obtains when the executions of all the implementations used by the measure to evaluate this principle are successful. More precisely, the evaluations of all the implementations under the principle are assumed to be at their maximum value (i.e. $v_{max}(imp)$ for a given imp) disregarding all others, which are typically set to 0, except in presence of dependencies. Thus we first introduce function $best_n$:

$$best_n(imp) = \begin{cases} v_{max}(imp) & \text{if } imp \in \text{desc}(n) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Then, the impact of a node n, noted impact(n) is defined as the ratio of the node's best score to the maximum achievable score according to the measure:

$$\text{impact}(\mathcal{M}, n) = \frac{score(\mathcal{M}, best_n, \text{FAIR})}{score(\mathcal{M}, best_{FAIR}, \text{FAIR})} \tag{8}$$

6 Experimental Results

Through these experiments, we show that digital resources may obtain very different FAIR assessment scores when using different FAIR evaluation engines. In addition, we show that our proposed framework can precisely document the coverage and the relative importance, for each tool, of both fine-grained FAIR indicators as well as global principles, thus providing insights for users and tool developers on possible evaluation biases. Additional material is available online[5].

6.1 Do FAIR Assessment Engines Reach Consensus on FAIRness ?

Table 1 reports FAIR assessments for a collection of 10 scientific digital resources with F-UJI and FAIR-Checker. In this selection we aim at covering diverse domains with different types of resources such as datasets, ontologies, software, or training material. These resources are exposed on the web through diverse modalities such as institutional open data platforms, community specific registries (bioinformatics tools, machine learning models), e-learning platforms, legacy websites or raw metadata.

The F-UJI scores, expressed in percentages, have been manually collected from the tool's web interface [14]. The FAIR-Checker scores have been collected through the tool's API [15]. Since FAIR-Checker only provides fine-grained scores as reported in section 4, we computed a global score as a percentage based on the maximum achievable score.

Table 1 shows a relatively good agreement between the two engines for the first 5 entries (50% of our collection) with a standard deviation lower than 10. The best agreement appears for very high or very low scores. However, the evaluations provided by F-UJI appear to be more fine-grained. The online

[5] `https://github.com/ICG4FAIR/ICG4FAIR`

Table 1: Multiple FAIR assessments, ranked by standard deviation.

Resource	F-UJI (%)	FAIR-Checker (%)	Std dev
Dataset (PANGAEA) [31]	91	91.70	0.49
Gene Ontology (OLS) [21]	18	16.70	0.92
Dataset (Harvard Dataverse) [23]	75	79.20	2.97
Dataset (Kaggle) [26]	60	70.80	7.64
Online course (Moodle) [28]	4	16.70	8.98
Dataset (Governmental platform) [22]	52	70.80	13.29
Dataset (WHO) [39]	27	50.00	16.26
Training material (TeSS) [36]	39	70.80	22.49
Bioinformatics tool (bio.tools) [6]	18	54.20	25.60
Dataset (RDF metadata) [33]	43	87.50	31.47

Table 2: Evaluating a bio.tools catalogue record.

	FAIR Score (%)	F (%)	A (%)	I (%)	R (%)
F-UJI	18.8	**35.7**	33	**0**	10
FAIR-Checker	54.2	**75**	50	**66.7**	16.7

course (Moodle) and Gene Ontology both obtain the FAIR-Checker minimal score (16.7%) but obtain different scores with F-UJI, 4% and 18% respectively, suggesting that F-UJI evaluation is more detailed. In addition, for the second half of our resource collection, the FAIR assessment scores begin to diverge with a standard deviation ranging from 13.29 to 31.47. Globally, we observe that FAIR-Checker provides higher scores compared to F-UJI.

For the last two entries the scores are very different with a standard deviation greater than 25. It is completely reasonable to wonder why the evaluation results are so different. Is it due to the way FAIR assessment engine retrieve metadata, as described in [37]? Is it due to the engine inner implementations? In the next paragraphs, we feed our model and compare FAIR-Checker and F-UJI, when evaluating a record of a bioinformatics tools catalogue.

Now we focus on how impact, granularity and coverage quantities can help in understanding divergent FAIR assessments. Table 2 reports very different results for the global FAIR assessment of the bio.tools record [6]. If we explore in more details each individual principle, we highlight that findability and interoperability scores greatly differ.

Table 3 shows how F-UJI and FAIR-Checker differ in terms of impact, granularity and coverage. We can see that reusability has a highest impact (41.67%) on the global assessment score compared to FAIR-Checker (25%). This could contribute to the explanation of the very low global FAIR assessment of the bio.tools record in F-UJI (Table 2) compared to FAIR-Checker. The findability of the bio.tools record is better scored in FAIR-Checker (75%) compared to F-UJI (35.7%). However, this result should be interpreted with caution due to a poor

Table 3: Comparing F-UJI and FAIR-Checker (FC)

	Impact		Granularity		Coverage	
	F-UJI	FC	F-UJI	FC	F-UJI	FC
F	29.17	33.33	1.25	**2**	100	**50**
A	12.5	16.67	2	1	66.67	50
I	**16.67**	**25**	1	1	100	100
R	**41.67**	**25**	1.25	1	100	100

coverage of F principles in FAIR-Checker (50%), compared to F-UJI (100%). In addition, despite a low coverage, we show that this findability principle has still the higher impact (33.33%) in the global assessment, which is questionable. Regarding the interoperability, we observe 0% for F-UJI and 66.7% for FAIR-Checker. Both engines have the same granularity (1) for an 100% coverage, meaning that the two engines, by design, do not agree on the indicators for interoperability, or that their implementation greatly vary.

6.2 Comparison of Measures Based on the Characteristic Quantities

We illustrate the use of the characteristic quantities of the measures introduced in section 5 to objectively highlight their salient features and some of their differences. All the seven measures considered in section 4 have been represented in the generic model following the methodology explained in the same section.

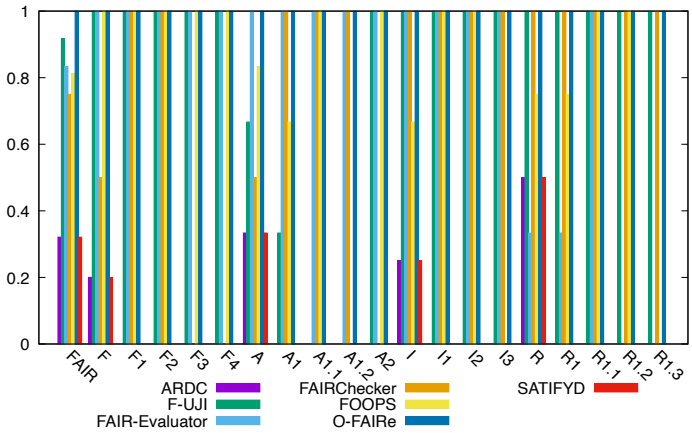


Fig. 2: Coverage rates.

Coverage Rates and Granularities The coverage rate and granularity of all the principles, for all the seven measures are shown in Figure 2 and Figure 3 respectively. Notice that concerning O'FAIRe, the coverage rate and the granularity do not consider that some of its indicators are not (yet) implemented. This does not change the coverage value since all principles have at least one indicator implemented. However, the granularity is slightly overestimated.

We highlight some elements of the analysis that can be made. First, the coverage rate of root FAIR is rarely equal to 1. This means that the majority of the measures we are studying do not cover one or several principles. Only O'FAIRe covers all the principles. Questionnaires ARDC and SATIFYD have a low coverage rate because they only consider principles F, A, I and R. For FOOPS!, the coverage rate of A is 0.83. Its value is 0.66 for A1 whereas one would expect less because the value is 1 for A1.1 while A1.2 is not covered at all. This is because of the presence of an indicator directly related to A1. We also observe that some sub-principles are not covered by each measure. By design, R1.2 and R1.3 are not covered in the FAIR Evaluation Service as well as F3 and F4 for FAIR-Checker. In addition some principles such as A1.2 and R1.3 are covered by only a few number of tools, which questions on the technical feasibility of their implementation.

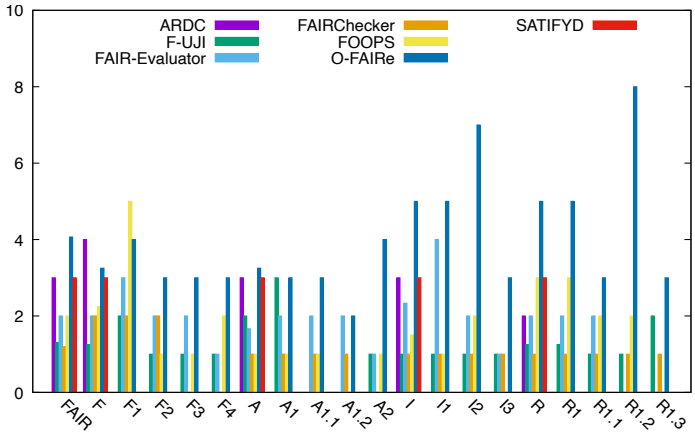


Fig. 3: Granularities

O'FAIRe gets the highest granularity at root FAIR, with high scores for I2 and R1.2 in particular. This is because it defines many indicators may be to address the great variety of vocabularies and meta-data present within the semantic artifacts it usually assess. As for FOOPS!, the high value of granularity of F1 shows an important care in providing indicators for this principle. Notice too that the granularity of R1 is higher than the granularities of R1.1, R1.2, R1.3. This is because several indicators have been directly attached to R1.

In fact, granularity and coverage rate are complementary and should sometimes be considered together before drawing a conclusion. For example, the promising granularity of F for ARDC could mean that F has been paid a lot of attention. However, the coverage rate of F for ARDC is quite low, meaning that there is no analysis according the sub-principles of F. Contrary to what one could expect, the analysis is not that precise.

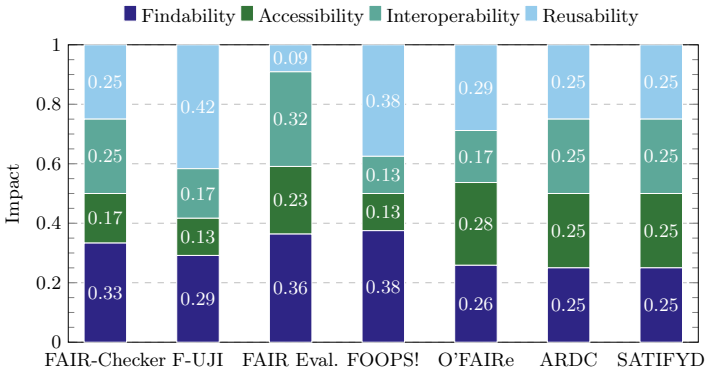


Fig. 4: Impact of the main FAIR Principles for different measures

Impacts We then compare the different measures according to the impact, i.e. the importance they give to each principle. Figure 4 illustrates the impact of the four main principles, but we could compare the measures in detail according to each sub-principles. Findability and reusability are the most important principles in general. Reusability even count as 42% of the maximum score for F-UJI. But unlike the other measures, it is of really low importance for FAIR Evaluation Service, which does not consider R1.2 and R1.3. Overall, these measures are not balanced, with a principle that is almost three times more important than another one for three of the five considered measure.

Apart from O'FAIRe, none of the measures really think about the importance to give to each indicator. This is understandable from the point of view of certain measures. For instance, Wilkinson *et al.* [41] insist on the fact that FAIRness measures should (only) act as an incentive to improve the FAIRness of digital resource, that "there is no intrinsic value in an evaluation score" and that we should not declare a resource FAIR or non-FAIR. However, even if the measures are only intended to push for improvement in the FAIRness of resources, it is a unfortunate that there is no order of priority. The SHARing Rewards and Credit (SHARC) Interest Group [12] and the Data Maturity Model Working Group [4], both from the Research Data Alliance, developed this idea that some criteria are more important than others by categorizing them as "useful", "important" or "essential".

Table 4: Impacts for FAIR-Checker, with and without dependencies.

	F	F1	F2	F3	F4	A	I	I1	I2	I3	R	R1	R1.1	R1.2	R1.3
noDep	0.33	0.16	0.16	0	0	0.16	0.25	0.08	0.08	0.08	0.25	0.25	0.08	0.08	0.08
Dep	**0.58**	0.16	**0.41**	0	0	0.16	**0.50**	**0.16**	**0.25**	0.08	**0.41**	**0.41**	0.08	0.08	**0.25**

Dependency Aware Analysis Our last experiment highlights the importance of taking into account dependencies. Those may stem from different codes checking a same property, or from full code delegations such as those expressed by FAIR-Checker and presented in Figure 1a. The first line in Table 4 shows impacts without dependencies. For the second one, to compute the impact of a node n, instead of systematically set to 0 the implementations that do not belong to desc(n), their values are set according to the dependencies. For example, when computing impact(F), implementation `iI1` is set to v_{max}(`iF2A`), which further increases the impact of F. Hence, the success of all the indicators belonging to desc(F) ensures to obtain not 33% but 58% of the maximum possible score. From a user point of view, such analysis is quite important, since it reveals that F is much more central to this measure than one might think at first glance.

7 Concluding Remarks

In this paper, we introduce a generic model and three computable and objective quantities aimed at more precisely interpreting FAIRness measures, comparing tools and possibly revealing evaluation biases. By adapting the hierarchy of principles, our framework could be repurposed for different evaluations, including IT security or energy footprint. Our experiments show that our framework i) contributes to explain different scores obtained by the same digital artifacts using different assessment engines and ii) facilitates the setup of comparative studies of various FAIRness metrics. Our approach is intended to be generic and to cover a large spectrum of FAIR assessment use cases. However, some of our choices induce some inaccuracies concerning granularity. Indeed, F-UJI divides its indicators into a new level of tests that would increase its granularity if considered. We provide experimental results on a limited set of resources. We are convinced that our generic model would benefit from larger scale experiments, with more FAIR assessment tools. As future works, we intend to more deeply analyze links between indicators or implementations, and to conduct larger scale experiments. This would require time, expertise, and would clearly be facilitated by involving tool development teams. We thus aim at contributing to collective initiatives tackling the challenges of harmonizing FAIR assessment frameworks.

Acknowledgments. This work is supported by the ANR DeKaloG (Decentralized Knowledge Graphs) project, ANR-19-CE23-0014, CE23 - Intelligence artificielle.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amdouni, E., Bouazzouni, S., Jonquet, C.: O'faire makes you an offer: metadata-based automatic fairness assessment for ontologies and semantic resources. International Journal of Metadata, Semantics and Ontologies **16**(1), 16–46 (2022)
2. Fair evaluation service, https://w3id.org/AmIFAIR [Accessed: 18 April 2024]
3. Australian Research Data Commons (ARDC): FAIR self assessment tool: Ardc online questionnaire (2022), https://ardc.edu.au/resource/fair-data-self-assessment-tool/ [Accessed: 22 April 2024]
4. Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., Stall, S.: The fair data maturity model: An approach to harmonise fair assessments. Data Science Journal **19**, 41–41 (2020)
5. Barker, M., Hong, N.P.C., Katz, D.S., Lamprecht, A.L., Martinez-Ortiz, C., Psomopoulos, F.E., Harrow, J., Castro, L.J., Gruenpeter, M., Martínez, P.A., Honeyman, T.: Introducing the fair principles for research software. Scientific Data **9** (2022), `https://api.semanticscholar.org/CorpusID:252878844`
6. Sample bioinformatics tool, https://bio.tools/bwa [Accessed: 22 April 2024]
7. Candela, L., Mangione, D., Pavone, G.: The fair assessment conundrum: Reflections on tools and metrics. Data Sci. J. **23**, 33 (2024), `https://api.semanticscholar.org/CorpusID:270073165`
8. Clarke, D.J., Wang, L., Jones, A., Wojciechowicz, M.L., Torre, D., Jagodnik, K.M., Jenkins, S.L., McQuilton, P., Flamholz, Z., Silverstein, M.C., et al.: Fairshake: toolkit to evaluate the fairness of research digital resources. Cell systems **9**(5), 417–421 (2019)
9. Corcho, Ó., Ekaputra, F.J., Heibi, I., Jonquet, C., Micsik, A., Peroni, S., Storti, E.: A maturity model for catalogues of semantic artefacts. Scientific Data **11** (2023), `https://api.semanticscholar.org/CorpusID:258615711`
10. Csiro misc questionnaire, https://web.archive.org/web/20210813120307/http://5stardata.csiro.au/ [Accessed: 11 July 2024]
11. Data Archiving and Networked Services (DANS): Satifyd online questionnaire (2019), https://satifyd.dans.knaw.nl/ [Accessed: 22 April 2024]
12. David, R., Mabile, L., Specht, A., Stryeck, S., Thomsen, M., Yahia, M., Jonquet, C., Dollé, L., Jacob, D., Bailo, D., et al.: Fairness literacy: The achilles' heel of applying fair principles. CODATA Data Science Journal **19**(32), 1–11 (2020)
13. Devaraju, A., Huber, R.: An automated solution for measuring the progress toward fair research data. Patterns **2**(11), 100370 (2021). `https://doi.org/https://doi.org/10.1016/j.patter.2021.100370`, `https://www.sciencedirect.com/science/article/pii/S2666389921002324`
14. F-uji misc tool, https://www.f-uji.net/index.php?action=test [Accessed: 16 April 2024]
15. Fair-checker tool, https://fair-checker.france-bioinformatique.fr [Accessed: 18 April 2024]
16. Foops! misc tool, https://foops.linkeddata.es/FAIR_validator.htm [Accessed: 18 April 2024]
17. Gaignard, A., Rosnet, T., De Lamotte, F., Lefort, V., Devignes, M.D.: Fair-checker: supporting digital resource findability and reuse with knowledge graphs and semantic web standards. Journal of Biomedical Semantics **14**(1), 1–14 (2023). `https://doi.org/10.1186/s13326-023-00289-5`

18. Garijo, D., Corcho, O., Poveda-Villalón, M.: FOOPS!: An ontology pitfall scanner for the fair principles. In: International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks. CEUR Workshop Proceedings, vol. 2980. CEUR-WS.org (2021), http://ceur-ws.org/Vol-2980/paper321.pdf
19. Peters-von Gehlen, K., Höck, H., Fast, A., Heydebreck, D., Lammert, A., Thiemann, H.: Recommendations for discipline-specific fairness evaluation derived from applying an ensemble of evaluation tools. Data Science Journal **21**, 7–7 (2022)
20. Genova, F., Aronsen, J.M., Beyan, O., Harrower, N., Holl, A., Hooft, R.W., Principe, P., Slavec, A., Jones, S.: Recommendations on FAIR metrics for EOSC. Publications Office of the European Union (2021)
21. Gene ontology, https://www.ebi.ac.uk/ols4/ontologies/go [Accessed: 22 April 2024]
22. Sample gouvernemental dataset, https://www.data.gouv.fr/en/datasets/donnees-relatives-a-lepidemie-de-covid-19-en-france-vue-densemble/ [Accessed: 22 April 2024]
23. Sample harvard dataset, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JGO6VI [Accessed: 22 April 2024]
24. Huerta, E.A., Blaiszik, B., Brinson, L., Bouchard, K.E., Diaz, D., Doglioni, C., Duarte, J.M., Emani, M.K., Foster, I.T., Fox, G., Harris, P.C., Heinrich, L., Jha, S., Katz, D.S., Kindratenko, V., Kirkpatrick, C.R., Lassila-Perini, K., Madduri, R.K., Neubauer, M.S., Psomopoulos, F.E., Roy, A., Rübel, O., Zhao, Z., Zhu, R.: Fair for ai: An interdisciplinary and international community building perspective. Scientific Data **10** (2022), https://api.semanticscholar.org/CorpusID:260201856
25. Jones, S., Grootveld, M.: How fair are your data? (Jul 2021). https://doi.org/10.5281/zenodo.5111307, https://doi.org/10.5281/zenodo.5111307
26. Sample kaggle dataset, https://www.kaggle.com/datasets/imdevskp/corona-virus-report [Accessed: 22 April 2024]
27. Krans, N., Ammar, A., Nymark, P., Willighagen, E., Bakker, M., Quik, J.: Fair assessment tools: evaluating use and performance. NanoImpact **27**, 100402 (2022)
28. Sample moodle course, https://moodle.polytechnique.fr/course/index.php?categoryid=1018 [Accessed: 22 April 2024]
29. Moser, M., Werheid, J., Hamann, T., Abdelrazeq, A., Schmitt, R.H.: Which fair are you? a detailed comparison of existing fair metrics in the context of research data management. In: Proceedings of the Conference on Research Data Infrastructure. vol. 1 (2023)
30. O'faire misc tool, https://agroportal.lirmm.fr/landscape#fairness_assessment [Accessed: 18 April 2024]
31. Sample pangaea dataset, http://doi.org/10.1594/PANGAEA.908011 [Accessed: 22 April 2024]
32. Poveda-Villalón, M., Espinoza-Arias, P., Garijo, D., Corcho, Ó.: Coming to terms with fair ontologies. In: International Conference Knowledge Engineering and Knowledge Management (2020), https://api.semanticscholar.org/CorpusID:225078634
33. Sample rdf metadata, https://data.rivm.nl/meta/srv/eng/rdf.metadata.get?uuid=1c0fcd57-1102-4620-9cfa-441e93ea5604&approved=true [Accessed: 22 April 2024]
34. Slamkov, D., Stojanov, V., Koteska, B., Mishev, A.: A comparison of data fairness evaluation tools. In: Budimac, Z. (ed.) Proceedings of the Ninth Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications, Novi Sad,

Serbia, September 11-14, 2022. CEUR Workshop Proceedings, vol. 3237. CEUR-WS.org (2022), https://ceur-ws.org/Vol-3237/paper-sla.pdf
35. Sun, C., Emonet, V., Dumontier, M.: A comprehensive comparison of automated fairness evaluation tools. In: 13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. pp. 44–53 (2022)
36. Sample training material, https://tess.elixir-europe.org/materials/make-your-research-fairer-with-quarto-github-and-zenodo [Accessed: 22 April 2024]
37. Van De Sompel, H., Soiland-Reyes, S.: Fair signposting: Exposing the topology of digital objects on the web. In: International FAIR Digital Objects Implementation Summit 2024. TIB Open Publishing (2024)
38. de Visser, C., Johansson, L.F., Kulkarni, P., Mei, H., Neerincx, P.B.T., van der Velde, K.J., Horvatovich, P., van Gool, A.J., Swertz, M.A., 't Hoen, P.A.C., Niehues, A.: Ten quick tips for building fair workflows. PLOS Computational Biology **19** (2023), https://api.semanticscholar.org/CorpusID:263224298
39. Sample who dataset, https://data.who.int/dashboards/covid19/data [Accessed: 22 April 2024]
40. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)
41. Wilkinson, M.D., Dumontier, M., Sansone, S.A., Bonino da Silva Santos, L.O., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M., et al.: Evaluating fair maturity through a scalable, automated, community-governed framework. Scientific data **6**(1), 174 (2019)
42. Wilkinson, M.D., Sansone, S.A., Marjan, G., Nordling, J., Dennis, R., Hecker, D.: FAIR Assessment Tools: Towards an "Apples to Apples" Comparisons (Jan 2023). https://doi.org/10.5281/zenodo.7463421, https://doi.org/10.5281/zenodo.7463421