A Comparison of Data FAIRness Evaluation Tools

Dejan Slamkov[1,†], Venko Stojanov[1,†], Bojana Koteska[1,*,†] and Anastas Mishev[1,†]

[1]Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, North Macedonia

Abstract

FAIR data principles represent a set of community-agreed guiding principles and practices for all researchers involved in the eScience ecosystem. The FAIR data principles were created to improve the reuse of data by making it findable, accessible, interoperable and reusable. The goal of these principles is to ensure that the inputs and outputs from computational analysis can be easily found and understood by data consumers, both humans and machines. Since the introduction of FAIR Data Principles in 2016, the interest in these principles has been constantly increasing and several research groups have started developing tools for evaluation of data FAIRness. In this paper, we aim to analyze the available online tools and checklists for data FAIRness evaluation and to provide tool comparison based on multiple features. Taking into account this analysis and tools advantages and disadvantages, we provide recommendations about the tools usage. A FAIRness practical evaluation is also conducted on seven data sets from different data repositories using the analysed tools. Findings show that there are no commonly accepted requirements evaluation of data FAIRness. The conclusions of this study could be used for further improvement of the FAIRness criteria design and making FAIR feasible in daily practice.

Keywords

Data FAIRness, open science, findability, reusability, interoperability, accessibility

1. Introduction

Today's exploitation of data shapes how we all live and function [1]. A growing number of electronics around us and on the Internet are allowing for the enormous growth of data [2]. International Data Center (IDC) predicts that by 2025 the total data produced would rise from 33 Zettabytes (ZB) in 2018 to 175 ZB [1]. If we look at planes, they produce around 2.5 billion Terabyte of data from the sensors mounted in the engines, per year [2]. It's like everywhere we turn to, we are surrounded by data.

As all things on the Internet, we would like the data to be easily discovered and consumed by users, just like websites on the Web. Therefore, research communities around the world have gathered to draft principles to improve the consumption of data on the Internet, thus the FAIR principles were born. FAIR is an abbreviation of the words Findable, Accessible, Interoperable

SQAMIA 2022: Workshop on Software Quality, Analysis, Monitoring, Improvement, and Applications, September 11–14, 2022, Novi Sad, Serbia

*Corresponding author.

†These authors contributed equally.

dejan.slamkov@students.finki.ukim.mk (D. Slamkov); venko.stojanov@students.finki.ukim.mk (V. Stojanov); bojana.koteska@finki.ukim.mk (B. Koteska); anastas.mishev@finki.ukim.mk (A. Mishev)

https://www.finki.ukim.mk/en/content/bojana-koteska-phd (B. Koteska); https://www.finki.ukim.mk/en/staff/anastas-mishev (A. Mishev)

0000-0001-6118-9044 (B. Koteska)

and Reusable. The FAIR principles are a guide on achieving FAIR data, not a set of rules to follow. With that being said, FAIR is not a standard, it does not define the how-to's, it is completely open to different interpretations and should not be used to assess the quality of data [3]. Instead, FAIR are guidelines for better experience with the data on the Internet, for both humans and machines. First formulation of the FAIR data vision was actually in 2014 and the primary goal was to optimise data sharing and reuse by humans and machines. In 2016, this initiative resulted in the first formal publication of FAIR principles, "The FAIR Guiding Principles for scientific data management and stewardship" by Wilkinson et al [4].

"Findable" yields ways for others to easily discover the data, e.g. using public repositories for data storage and assigning Data Object Identifiers (DOIs) for citation. "Accessible" allows for maximum availability of the data i.e. designs the access (and restrictions) to the data using the Internet protocols (FTP, HTTPS). "Interoperable" will make sure that the data is easily integrated with other data and can easily be consumed by both machines and humans, meaning multiple file formats for the machines and usage of widely used languages for the humans. "Reusable" provides for easier usage and understanding of the data by other researchers by requesting metadata and documentation [5]. One common misconception about the FAIR principles is that: "FAIR data means open data". The FAIR principles allow (and encourage) licences, which can restrict the access to the data.

Open Access Infrastructure for Research in Europe (OpenAIRE) [6] which is one of the leading projects for open science in Europe and infrastructure for open scholarly and scientific communication, also promotes and supports FAIR principles. According to OpenAIRE, FAIR principles describe the organization of the research outputs so they can be more easily accessed, understood, exchanged and reused. In more details, **Findable** requires data with persistent identifier, rich metadata, searchable and discoverable online; **Accesible** means data retrievable online using standardised protocols and restrictions if necessary; **Interoperable** recommends using common formats and standards and controlled vocabularies and **Reusable** imposes well-documented data with clear licence and provenance information (Fig. 1).

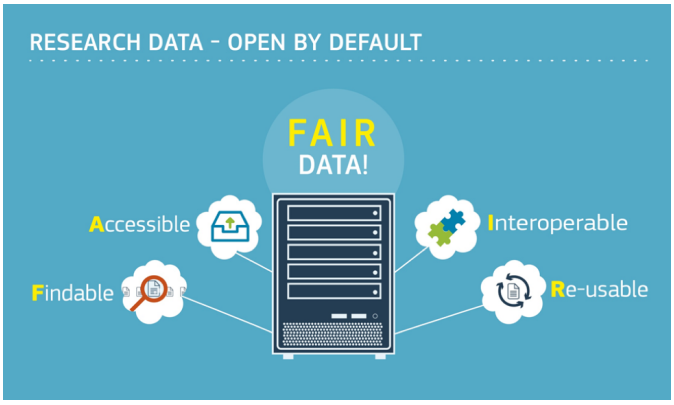

Figure 1: FAIR principles by OpenAIRE [6].

In this paper, we aim to provide insights at specific tools for FAIR data assessment and analyze their capabilities. In Section 2 we cover some related work. Section 3 provides descriptions about

fair assessment tools characteristics, with the exception of the last subsection, which draws parallels between them based on multiple relevant characteristics. In Section 4 we evaluate open data sets using the tools we addressed in the previous section. At last, we draw the conclusion of this paper by revealing the advantages and disadvantages of the tools, with respect to their "FAIRness" evaluation and characteristics.

2. Related work

The FAIR data principles requirements and evaluation are relatively new topics that started to be researched in 2016 and have been rapidly increasing. With the need data to be easily discovered and consumed by users, researches have begun to access and evaluate data FAIRness.

Camilla Hertil Lindelöw et al. [7] talk about the Swedish Government allocating parallel assignments to the Swedish National Library in order to develop criteria and a mechanism for assessing how well research data and scholarly publications created in Swedish organizations comply with the FAIR principles. The discussion describes recommendations and the possibilities and setbacks that they have identified during the work, focusing primarily on evaluation at a national level. Thompson at al. [8] describe tools that aid the FAIR process, from FAIR data management planning to FAIR data creation, publication, evaluation and (re)use, revealing that there are a lot of ongoing efforts that contribute to the goal of making FAIR a reality. In [9], the authors describe a FAIR framework and execute compliance tests with the FAIR metrics. They demonstrate its usage in some commonly used repositories and provide feedback where semi-automated evaluations are performed. It's revealed that the distinction between manual and automatic assessment shows that automatic assessments are usually more rigorous, resulting in lower FAIRness scores, though more precise. In [10], the authors briefly outline the different kinds of FAIRness evaluations describing the pros and cons of each kind and provide guidelines on how FAIRness evaluations can be used and interpreted. They define discrete-answer questionnaire-based evaluations, open-answer questionnaire-based evaluations and semi-automated evaluation, concluding that evaluations should be assessed not at the overall FAIRness level, but at the maturity indicator level. Bishop et al.[11] explore how the FAIR principles can be measured for re-use from a consumer perspective, stating that some FAIR principles can be subjectively automated more than others and that requires more qualitative, subjective measures for automation. They provide recommendations to create context aware questionnaires to evaluate the FAIR principles in a way that captures the FAIR-ness from the perspective of data re-user/consumer. Mons et al. [12] discuss how the data will stay FAIR in the ever emerging cloud world. They explore the public and commercial domain of the cloud sphere and what they are willing to offer for open and FAIR data across cloud platforms. The FAIR principles are revised from the perspective of the European Open Science Cloud. Madduri et al. [13] examine tools designed to help implement complex "big data" computations in ways that allow the code and associated data to be FAIR. To highlight the usage of the tools, the authors present a case study on the implementation of a multi-stage DNase Hypersensitivity (DHSs) sequencing data analysis that retrieves massive data sets from a public repository and uses a combination of parallel cloud and workstation computing to identify binding sites of candidate transcription factors.

3. FAIR Data Assessment Tools and Analysis

In this Section we provide descriptions and comparison of multiple tools for data FAIRness evaluation.

3.1. Tools Descriptions

3.1.1. ARDC's tool

The Australian Research Data Commons (ARDC) is an organization that has been an active advocate for the adaptation of the FAIR principles. It provides multiple useful resources to the international research community to ensure the usage of best practices in their research, one of them being a FAIR data self assessment tool which provides a score describing the "FAIRness" of the data. It is important to note that the tool is made with the interpretation of the FAIR principles and it is made to trigger thinking and discussion around potential approaches of making data more FAIR [14]. The tool is hosted on the web site of ARDC [15]. It is designed in a form of a survey with 12 questions which are formulated to quantify the intensity of each FAIR category. It has 4 sections, titled by the four FAIR principles. There are single choice questions and questions in Yes/No format. The answers are evaluated based on previous answers for consistency. Every question essentially traces back to a certain FAIR principle. By answering a question, the bar on the bottom of the corresponding section fills up depending on the "strength" of the provided answer, i.e. it's measuring the presence of a FAIR principle. On the bottom (and top) the "Total across F.A.I.R" bar also fills up depending on the section score bars, representing the total "FAIRness" score of the data. The tool captures the FAIR essence. It addresses the FAIR principles in great detail and provides additional information on "FAIR terms". It is very concise and to the point, self-explanatory and has a nice design and layout. It does not disclose the algorithm behind the scoring system, but the pattern is easily noticeable after a few tries. It is easily accessible and needs no form of log-in.

3.1.2. SATIFYD

Data Archiving and Networked Services (DANS) urges researchers to make their digital research data and related outputs FAIR [16]. To assist the process, a FAIR data self-assessment tool called SATIFYD (Self-Assessment Tool to Improve the FAIRness of Your Dataset) was created. It is intended primarily to evaluate datasets that will be published on EASY, which is an online archiving system for depositing and reusing research data [17]. SATIFYD is in the form of a 12-question questionnaire divided into 4 sections, each section having questions centered around a certain FAIR principle. Some of the addressing aspects of the FAIR principles are guaranteed by EASY. For example, DOI requirement is omitted from the questions, thus meaning that SATIFYD relies on the repositories's ability to ensure some of the principles. Some questions are directly linked to services offered by EASY, and there is no work-around, meaning that the overall FAIR score would suffer if the data is not published on EASY. Also, some questions are repeated in different sections, since they capture the nature of several FAIR principles [16]. If the score is not perfect (100%), it offers guidance on how to lift it with pressing the "Want to improve?" button on that section. The consistency of the answers is tested by referring to previous question's

answers and if anything does not add up, a pop-up appears to expose the inconsistencies. It does not disclose the algorithm behind the scoring system. This tool also includes a feature for getting printed reports on the answers, together with the tips on improving.

3.1.3. CSIRO's tool

CSIRO's (Australia's Commonwealth Science and Industrial Research Organization) initiative OzNome develops tools and methods aimed at providing access to self-organizing, reliable, organized, and well-governed data ecosystems [18]. In that spirit, they developed a 5 star FAIR data self assessment tool, which is publicly hosted on their website [19]. The tool embodies the four FAIR categories, plus adding "Trusted" to the bunch, which is trying to determine if the data keeps records on how it's been used, by whom, and how many times. The tool is in the form of a survey with questions each yielding single-choice answers. Every FAIR principle is captured by a series of different questions which are designed to allow data rating according to its current state. The end result is a 5 star rating for each category for the data, showing the compliance to the 5 categories. The number of stars from the respective categories are adjusting with every answered question. The survey kicks off with the requirement for information about the name (title) and URL of the data, for which you can provide false information if none other information is available at the moment. These fields don't affect the score. The consistency of the answers is not checked and it does not provide any additional explanatory information on the site about the concepts covered by the tool. On the other hand, it discloses the rating scheme [20].

3.1.4. EUDAT checklist

The European Association of Databases for Education and Training (EUDAT) Collaborative Data Infrastructure (CDI) project holds an infrastructure for integrated data services and resources supporting research in Europe [21]. The CDI provides a common infrastructure that enables data management across European research communities, allowing researchers from any research discipline to preserve, find, access and process data in a trusted environment. The EUDAT checklist was developed in order to help the researchers to test the "FAIRness" of their data [22]. The checklist is not a self-assessment tool, but rather a handy reference sheet that can be printed out for a quick check on the FAIRness of the data. It contains 4 sections, each yielding a brief summary of a FAIR principle and 4 statements associated with it. For the purpose of this paper, the final score shall be calculated as the percentage of the checked off answers. The questionnaire is available in two versions, one as a more simple checklist, the other as a data flyer with better design and some colors. This checklist is also supported by the OpenAIRE project [6].

3.1.5. RDA checklist

Research Data Alliance (RDA) [23] is composed of multiple organizations from around the world with the purpose of strengthening the social and technological bridges to enable open data sharing and reuse. To stimulate data sharing, an interdisciplinary scientific interest group called SHAring Reward & Credit (SHARC) was set up by RDA. The primary aim of SHARC is to

find ways to promote the value of the data sharing process and find ways of crediting those who comply. With these priorities in mind, SHARC developed a simplified evaluation grid which employs criteria against researcher's data to determine the presence of the FAIR principles [24]. The simplified grid is intended for use by researchers who produce and/or use data. It includes 4 sections representing the FAIR principles. Each section contains several questions, which for the purpose of this paper will be answered only by YES or NO. The questions from the sections are in the form of a decision tree, indicating that there are preconditions for some questions. For example, to answer: "Unique, global, persistent ID?", first the data has to be indexed, i.e. the question "Indexed identifier?" must be answered with "YES". For the purpose of this paper, the final score is the percentage of "YES" answered questions. The results of the grid are to be used for appreciating the researcher's practice and to spark discussion to keep the data management life cycle more "FAIR", but not for comprehensive data FAIRness assessment.

3.2. Tools Analysis

3.2.1. Selected Features

After a detailed study of the tools/checklists described in Subsection 3.1, we selected eight features (criteria) to make a comparison: **1.Automated evaluation** - Does not require manpower for calculating the final score; **2.Disclosing the rating system** - The algorithm behind the calculation of the final points is available to the user; **3.Additional explanatory information** - Provides all the needed info about the FAIR terms mentioned in the questions; **4.Bound to a repository** - It relies on features made available by a certain repository to impose certain FAIR principles; **5.Guidance to improve "FAIRness"** - Provides tips and information on how to comply more heavily to the FAIR principles; **6.Printed report** - Provides a printed report of all the answers; **7.Goes beyond FAIR** - The questions cover principles outside of FAIR; **8.Checklist** - The questions are only in a Yes/No format.

3.2.2. Comparison by Features and Recommendations

Table 1 presents the comparison of FAIR data assessment tools. First column denotes the tool name, while first row contains the eight features as explained in Subsection 3.2.1. The symbol "✓" represents that the tool "posses" the selected feature. For example, the tool ARDC has Additional explanatory information (feature 3).

Table 1
Fair Data Assessment Tools Comparison Matrix.

Feature/ Tool	1	2	3	4	5	6	7	8
ARDC	✓	✗	✓	✗	✗	✗	✗	✗
SATIFYD	✓	✗	✓	✓	✓	✓	✗	✗
CSIRO	✓	✓	✗	✗	✗	✗	✓	✗
EUDAT checklist	✗	✓	✗	✗	✗	✓	✗	✓
RDA checklist	✗	✓	✗	✗	✗	✓	✗	✓

Based on the tools analysis we provide the following recommendations:

- If EASY is not the chosen repository where the data will be published then SATIFYD tool should be avoided, because the overall score suffers greatly (as we will see in the next Section);
- The checklists are more vague on how the FAIR principles are truly obtained. On the other hand, the non-checklists provide a palette of specific answers based on FAIR principles implementations, which can affect the final score in different ways;
- ARDC and SATIFYD provide an abundance of additional information about the FAIR terms and specifically cite their sources, so these tools would be great for researchers that are not that familiar with the FAIR principles or don't know where to look for information on them;
- CSIRO, EUDAT and RDA disclose their algorithms for researchers and they can give feedback to the creators and build their own tools on top of these algorithms;
- If EASY is the repository of choice, then SATIFYD is definitely the right choice to assess the FAIRness. It provides unique features that other tools lack, e.g. guidance for improvement. On the other hand, if the data is not/will not be published on EASY, then ARDC or CSIRO should be the go-tos if the researcher is not in need of a printed report, otherwise the checklists would do the job.

4. Evaluation of data FAIRness

4.1. Datasets

The evaluation of the data FAIRness was performed on seven open datasets chosen from the following repositories: **PANGAEA** - Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research and has been selected as one of the 6 data repositories that provides their expertise in testing practical solutions to enhance the FAIRness of data [25]; **PhysioNet** - repository of freely-available medical research data, managed by the MIT Laboratory for Computational Physiology [26]; **DRYAD** - international open-access repository of research data, especially data underlying scientific and medical publications. Dryad is a resource that makes research data discoverable, freely reusable, and citable [27]; **EASY** - online archiving system of the creators of the tool for FAIRness data evaluation SATIFYD, Data Archiving and Networked Services (DANS). EASY offers access to thousands of datasets in the humanities, the social sciences and other disciplines [17]; **Dataverse** - an open source data repository framework used by individual researchers, archives, academic institutions and publishers around the world to share, find, cite, and preserve research data [28]; **datagovmk** - contains datasets from different institutions from North Macedonia [29]; and **Arctic Data Center** - primary data and software repository for the Arctic section of NSF Polar Programs which is said to be in "large degree already compliant with the FAIR principles" [30].

Table 2 presents the metadata for the seven datasets: Dataset name, Repository, Short Desrciption, Dataset Size, Creators and Year of publishing.

Table 2
Datasets metadata.

Name	Repository	Description	Size	Creators	Year
Fish survey during July-August 2016 at a Bahamian coral reef [31]	PANGAEA	This dataset includes a complete visual census of fish underwater at Cape Eleuthera, the Bahamas. The data are divided into the four sites (Tunnel Rock, Cathedral, Some2C and Ike's Reef) and further into species.	84 KB	Zhu, Yiou; Newman, Steven P; Reid, William D K; Polunin, Nicholas V C	2019
PTB-XL - large publicly available electrocardiography dataset [32]	PhysioNet	The PTB-XL ECG dataset comprises 21837 clinical 12-lead ECG (Electrocardiography) records of 10 seconds length from 18885 patients, where 52% are male and 48% are female with ages covering the whole range from 0 to 95 years.	3 GB	Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, Tobias Schaeffter	2020
Upper Columbia River Steelhead Capture-Recapture-Recovery data (2008-2018) [33]	DRYAD	The dataset is composed of ESA -listed steelhead trout that were tagged (n = 78,409) and subsequently exposed to predation during smolt out-migration through multiple river reaches.	17 MB	Payton Quinn Hostetter Nathaniel	2020
Landslide inventory of the 2018 monsoon rainfall in Kerala, India [34]	EASY	The dataset contains a complete landslide inventory for the 2018 Monsoon landslide event in the state of Kerala, India collected with the purpose of analyzing the relationship between the intensity of the trigger (e.g. rainfall, earthquake) and the density of the landslides in a given area.	14.5 MB	Westen, Dr C.J. van	2020
Water sources in the Syrian Desert [35]	Dataverse	The dataset provides the location of 2236 water sources in the Syrian Desert that were originally printed on Soviet topographic maps in 1980. These consist of 853 pools / reservoirs / cisterns, 1061 small wells, 119 large wells and 203 springs.	7.21 MB	Seland, Eivind Heldaas	2019
Unnamed resource [36]	data.gov.mk	This dataset contains data about location, names, telephone numbers etc. of court legal translators in North Macedonia.	400.5 KB	Ministry of Justice, North Macedonia	2018
Temperature measurements from boreholes along the Alaskan Pipeline Project, 2015-2016 [37]	Arctic Data Center	This dataset contains data about temperatures measurements of boreholes that drilled over 120 boreholes in a transect between 2009 and 2012 from the Alaska/Canada border.	29.665 KB	Vladimir Romanovsky, Alexander Kholodov, William Cable, Lily Cohen, Santosh Panda	2017

4.2. Results

We evaluated the seven datasets using each of the five tools/checklists (ARDC, SATIFYD, CSIRO, EUDAT checklist and RDA checklist). The evaluation results for the four FAIR principles are presented in Table 3. F stands for findable, A for accessible, I for interoperable, R for reusable

and FAIR is the total score obtained as average value of the four principles. For the CSIRO tool there is only result for the total FAIR score because this tool provides visual 5-star output for each principle.

Table 3
FAIR evaluation matrix.

Tool/ Repository	ARDC	SATIFYD	CSIRO	EUDAT	RDA
PANGAEA	F = 82.35% A = 70.00% I = 62.50% R = 71.43% **FAIR = 74.55**%	F = 33.00% A = 55.00% I = 50.00% R = 41.00% FAIR = 45.00%	FAIR = 55.00%	F = 75.00% A = 75.00% I = 50.00% R = 75.00% FAIR = 68.75%	F = 62.50 % A = 33.33% I = 0.00% R = 40.00% FAIR = 44.44%
PhysioNet	F = 82.35% A = 80.00% I = 37.50% R = 85.71% FAIR = 74.96%	F = 67.00% A = 55.00% I = 67.00% R = 60.00% FAIR = 62.00%	FAIR = 51.20%	F = 100.00% A = 75.00% I = 50.00% R = 75.00% **FAIR = 75.00**%	F = 75.00 % A = 66.67% I = 50.00% R = 60.00% FAIR = 66.67%
DRYAD	F = 76.47% A = 70.00% I = 37.50% R = 71.43% **FAIR = 66.82**%	F = 38.00% A = 55.00% I = 58.00% R = 41.00% FAIR = 48.00%	FAIR = 50.00%	F = 75.00% A = 75.00% I = 25.00% R = 75.00% FAIR = 62.50%	F = 75.00 % A = 66.67% I = 50.00% R = 40.00% FAIR = 61.11%
EASY	F = 88.23% A = 70.00% I = 62.50% R = 100.00%, FAIR = 84.35%	F = 67.00% A = 55.00% I = 58.00% R = 87.00% FAIR = 67.00%	FAIR = 56.80%	F = 100.00% A = 75.00% I = 75.00% R = 100.00% **FAIR = 87.50**%	F = 75.00% A = 33.30% I = 50.00% R = 40.00% FAIR = 55.56%
Dataverse	F = 88.23% A = 70.00% I = 62.50% R = 100.00% FAIR = 84.35%	F = 78.00% A = 55.00% I = 58.00% R = 93.00% FAIR=71.00%	FAIR = 76.80%	F = 100.00% A = 75.00% I = 75.00% R = 100% **FAIR = 87.50**%	F = 75.00% A = 33.30% I = 50.00% R = 80.00% FAIR = 66.67%
data.gov.mk	F = 47.06% A = 70.00% I = 12.50% R = 0.00% FAIR = 32.39%	F = 16.00% A = 5.00% I = 8.00% R = 6.00% FAIR = 9.00%	FAIR = 37.60%	F = 75.00% A = 75.00% I = 25.00% R = 25.00% **FAIR = 50.00**%	F = 37.50% A = 33.33% I = 0.00% R = 20.00% FAIR = 27.78%
Arctic Data Center	F = 88.23% A = 80.00% I = 37.50% R = 100.00% FAIR = 80.60%	F = 56.00% A = 55.00% I = 58.00% R = 74.00% FAIR = 61.00%	FAIR = 58.40%	F = 100.00% A = 75.00% I = 75.00% R = 100.00% **FAIR = 87.50**%	F = 75.00 % A = 33.33% I = 50.00% R = 40.00% FAIR = 55.56%

The general overview shows that the manual checklist EUDAT gives the highest scores, which is a direct consequence of the vagueness of the questions in the questionnaire, thus shunning away from testing for concrete solutions/standards for certain principles.

On the other hand, RDA gives relatively lower scores than EUDAT, even though it has the

same problem as EUDAT: vagueness, but with the opposite effect. Some of the questions simply cannot be answered and thereby left unchecked, which decreases the overall score. For example: “Do the data reuse control and data sharing arrangements meet the data protection and "local/national ethics requirements?”. This question requires knowledge from the law/moral standards, but it is not clear whose law/moral standards should be examined (Which country’s?). The dataset is open and available for anyone in the world, thus this question becomes hard to answer (especially in YES/NO format).

On the non-manual side, it’s observable the ARDC is the most “generous one”. This happens because ARDC is a little loose on the “R” side of FAIR, compared to SATIFYD and CSIRO, who are more strict and demand the implementations of several standards to increase the reusability of the dataset.

SATIFYD scores are lower because of the “EASY nature” of the tool. For example: it demands file formats that are specifically preffered by EASY and if those are not provided, the score suffers, even though these “non-preferred” formats are widely used and standardized.

5. Conclusion

In 2016 a group of researchers published a measurable set of principles for the academia and industry known as FAIR Data Principles. The primary intent was to develop a guideline for researches who want to enhance the reusability of their data. However, these principles set up a lot of challenges that need to be addressed. To understand the current research in the field of FAIR principles, we have analyzed the provided FAIR evaluation options of five tools and identified main characteristics and differences. We also perform FAIR evaluation of seven datasets from different data repositories using the five tools. The findings have shown that this topic is still in its early phase. The results show that the FAIR requirements are still not unified. For some datasets, we obtained scores that differ by more than 30%. It indicates that tools are somehow designed for a specific data repository. We can conclude that there is still no agreement for the universally accepted requirements for data FAIRness evaluation tools. This study contributes to the theory by analyzing the evaluation options and requirements for data FAIRness and by providing a guidance for tool selection and improvement.

Acknowledgments

This work was supported in part by the European Union’s Horizon 2020 research and innovation programme, project National Initiatives for Open Science - Europe, NI4OS-Europe, [857645] and by the Faculty of Computer Science and Engineering, Skopje, North Macedonia.

References

[1] D. Reinsel, J. Gantz, J. Rydning, The digitization of the world from edge to core, IDC White Paper (2018).

[2] M. Ghotkar, P. Rokde, Big data: How it is generated and its importance, IOSR Journal of Computer Engineering (2016).

[3] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, et al., Fair principles: interpretations and implementation considerations, 2020.
[4] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016).
[5] K. K. Hansen, M. Buss, L. S. Haahr, A fairy tale: A fake story in a trustworthy guide to the fair principles for research data (2018).
[6] OpenAIRE, Openaire, https://www.openaire.eu/, 2020. (Accessed on 05/24/2020).
[7] C. Lindelöw, Fair already? principles of reusability and research output–evaluation at a national level (2019).
[8] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos, L. O. B. da Silva Santos, Making fair easy with fair tools: From creolization to convergence, Data Intelligence (2020) 87–95.
[9] M. D. Wilkinson, M. Dumontier, S.-A. Sansone, L. O. B. da Silva Santos, M. Prieto, D. Batista, P. McQuilton, T. Kuhn, P. Rocca-Serra, M. Crosas, et al., Evaluating fair maturity through a scalable, automated, community-governed framework, Scientific data 6 (2019) 1–12.
[10] R. de Miranda Azevedo, M. Dumontier, Considerations for the conduction and interpretation of fairness evaluations, Data Intelligence (2020) 285–292.
[11] B. W. Bishop, C. Hank, Measuring fair principles to inform fitness for use (2018).
[12] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, M. D. Wilkinson, Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud, Information Services & Use 37 (2017) 49–56.
[13] R. Madduri, K. Chard, M. D'Arcy, S. C. Jung, A. Rodriguez, D. Sulakhe, E. Deutsch, C. Funk, B. Heavner, M. Richards, et al., Reproducible big data science: a case study in continuous fairness, PloS one 14 (2019).
[14] Australian Research Data Commons (ARDC), Australian Research Data Commons (ARDC), https://ardc.edu.au/, 2020. [Online; accessed 18-May-2020].
[15] Australian Research Data Commons (ARDC), Fair self assessment tool, https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/, 2020. [Online; accessed 18-May-2020].
[16] Data Archiving and Networked Services (DANS) , Organisation and policy, https://dans.knaw.nl, 2020. [Online; accessed 18-May-2020].
[17] Data Archiving and Networked Services (DANS) , EASY, https://easy.dans.knaw.nl/ui/home, 2020. [Online; accessed 18-May-2020].
[18] Australia's Commonwealth Science and Industrial Research Organization , OzNome Initiative, https://research.csiro.au/oznome/, 2020. [Online; accessed 18-May-2020].
[19] Australia's Commonwealth Science and Industrial Research Organization , CSIRO, http://oznome.csiro.au/5star/, 2020. [Online; accessed 18-May-2020].
[20] Australia's Commonwealth Science and Industrial Research Organization , CSIRO Rating Scheme, https://confluence.csiro.au/display/OZNOME/Data+ratings, 2020. [Online; accessed 18-May-2020].
[21] EUDAT, EUDAT Collaborative Data Infrastructure , https://eudat.eu/eudat-cdi, 2020. [Online; accessed 18-May-2020].
[22] EUDAT, How FAIR are your data?, https://zenodo.org/record/3405141, 2020. [Online;

accessed 18-May-2020].
[23] Research Data Alliance (RDA), About RDA, https://www.rd-alliance.org/about-rda, 2020. [Online; accessed 18-May-2020].
[24] Research Data Alliance (RDA), Data Sharing Evaluation to Trigger Crediting/Rewarding Processes, https://zenodo.org/record/2551500, 2020. [Online; accessed 18-May-2020].
[25] World Data Center PANGAEA, PANGAEA, https://www.pangaea.de/about/, 2020. [Online; accessed 18-May-2020].
[26] MIT Laboratory for Computational Physiology, PhysioNet, https://physionet.org/, 2020. [Online; accessed 18-May-2020].
[27] Dryad, Dryad Digital Repository, https://datadryad.org/stash/our_mission, 2020. [Online; accessed 18-May-2020].
[28] Harvard's Institute for Quantitative Social Science (IQSS) et al., Dataverse, https://dataverse.org/about, 2020. [Online; accessed 18-May-2020].
[29] Ministry of Information Society and Administration, North Macedonia, datagovmk, http://data.gov.mk/en/, 2020. [Online; accessed 18-May-2020].
[30] Arctic Data Center, Arctic Data Center repository, https://arcticdata.io/about/, 2020. [Online; accessed 18-May-2020].
[31] Y. Zhu, S. P. Newman, W. D. K. Reid, N. V. C. Polunin, Fish survey (total length and count) during July-August 2016 at a Bahamian coral reef, PANGAEA, 2019. URL: https://doi.org/10.1594/PANGAEA.898359. doi:10.1594/PANGAEA.898359, in: Zhu, Y et al. (2019): Fish survey (total length and count) and carbon and nitrogen stable isotope ratios of sampled fish during July-August 2016 at a Bahamian coral reef (Cape Eleuthera). PANGAEA, https://doi.org/10.1594/PANGAEA.898361.
[32] Wagner, P et al., PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1), PhysioNet, https://doi.org/10.13026/x4td-x982, 2020. [Online; accessed 18-May-2020].
[33] Quinn, P et al., Upper Columbia River Steelhead Capture-Recapture-Recovery data (2008-2018) , v4, Dryad Dataset, https://datadryad.org/stash/dataset/doi:10.5061/dryad.k98sf7m3r, 2020. [Online; accessed 18-May-2020].
[34] Westen, Dr C.J. van, Landslide inventory of the 2018 monsoon rainfall in Kerala, India, https://doi.org/10.17026/dans-x6c-y7x2, 2020. [Online; accessed 18-May-2020].
[35] E. H. Seland, Water sources in the Syrian Desert, 2019. URL: https://doi.org/10.18710/CEY9QR. doi:10.18710/CEY9QR.
[36] Ministry of Justice, Court legal translators in North Macedonia, http://www.data.gov.mk/en/dataset/cydckn-npebedybahn/resource/f6546a4d-e1a9-4f20-b0dd-d0565776040a, 2018. [Online; accessed 18-May-2020].
[37] Romanovsky, V et al, Temperature measurements from boreholes along the Alaskan Pipeline Project, 2015-2016, https://arcticdata.io/catalog/view/doi:10.18739/A2GM81P42, 2017. [Online; accessed 18-May-2020].