Recommendations for discipline-specific FAIRness evaluation derived from applying an ensemble of evaluation tools

Karsten Peters-von Gehlen*,[1], Heinke Höck*, Andrej Fast*, Daniel Heydebreck*, Andrea Lammert*, Hannes Thiemann*

*Deutsches Klimarechenzentrum GmbH,

Bundesstr. 45a, D-20146 Hamburg, Germany

[1]contact author: Karsten Peters-von Gehlen, peters@dkrz.de

currently under peer-review with Data Science Journal, submitted on 3 September 2021

Authors' Contributions

KPvG lead the process leading to the results presented in this paper in terms of conceiving the analysis methodology (together with HH) and writing of the manuscript. All other authors contributed substantially to the interpretation of work, revisited it critically for intellectual content, provided final approval for the work to be submitted, agreed to be accountable for the content of the study and agreed to be named in the author list.

Abstract

From a research data repositories' perspective, offering research data management services in-line with the FAIR principles is becoming increasingly important. However, there exists no globally established and trusted approach to evaluate FAIRness to date. Here, we apply five different available FAIRness evaluation approaches to selected data

archived in the World Data Center for Climate (WDCC). Two approaches are purely automatic, two approaches are purely manual and one approach applies a hybrid method (manual and automatic combined).

The results of our evaluation show an overall mean FAIR score of WDCC-archived (meta)data of 0.67 of 1., with a range of 0.5 to 0.88. Manual approaches show higher scores than automated ones and the hybrid approach shows the highest score. Computed statistics indicate that the test approaches show an overall good agreement at the data collection level.

We find that while neither one of the five valuation approaches is fully fit-for-purpose to evaluate (discipline-specific) FAIRness, all have their individual strengths. Specifically, manual approaches capture contextual aspects of FAIRness relevant for reuse, whereas automated approaches focus on the strictly standardized aspects of machine actionability. Correspondingly, the hybrid method combines the advantages and eliminates the deficiencies of manual and automatic evaluation approaches.

Based on our results, we recommend future FAIRness evaluation tools to be based on a mature hybrid approach. Especially the design and adoption of the discipline-specific aspects of FAIRness will have to be conducted in concerted community efforts.

1 Introduction

Since their original publication, the FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson et al., 2016) have initiated an advancement of research data management practices and requirements at an unprecendented pace. What the FAIR principles entail is essentially a formalization of what one would generally understand as the data management aspects of good scientific practice (Kruk, 2013), i.e. that digital objects forming the foundation of research results should be available to the global community in order to facilitate the validation of scientific results and enable broad reuse of scientific data.

Specifically, the FAIR principles have entered the day-to-day workflow of researchers, because funders and publishers more often than not require project data underlying scientific publications be managed, archived and made available to the scientific community in-line with the FAIR principles. Consequently, research data repositories and archives can offer the researchers a corresponding service if data curation practice in-line with the FAIR principles can be trustfully demonstrated and communicated. Indeed, current efforts to align the CoreTrustSeal[1] certification (Dillo & de Leeuw, 2018) with the FAIR principles are leveling the path in that regard (L'Hours et al., 2020) .

[1] https://www.coretrustseal.org

To date however, there exists no standardized and globally accepted procedure to truth- and trustfully evaluate the FAIRness of a research data repositories' (meta)data holdings and its data curation approach. Although recommendations regarding the metrics to be considered in FAIR evaluations have been recently published (Bahim et al., 2020; Genova et al., 2021), the lack of global agreement on and adoption of discipline-specific FAIR-ness criteria requires concerted community effort and remains a challenge (Wilkinson et al., 2019; Genova et al., 2021).

This does not mean that the development of FAIRness evaluation tools has not flourished. On the contrary, a plethora of tools - manual and automated as well as comprehensive and less comprehensive ones - has been and is continuously developed and is openly available for evaluating archived (meta)data (Bahim, Dekkers & Wyns, 2019). From the perspective of a repository operator aiming for FAIRness evaluation, it is however not evident which tool to choose from, because thorough evaluation of the fitness-for-purpose of the tools is not available.

In this study, we aim to close this knowledge gap by applying an ensemble of five different FAIRness evaluation tools to selected (meta)data archived in the World Data Center for Climate (WDCC)[2], which is hosted at the German Climate Computing Center (DKRZ)[3] in Hamburg, Germany. The WDCC is a CoreTrustSeal certified domain-specific archive for climate science, with a focus on ensuring the long-term reusability of climate simulation data and climate related data products. In earlier work, a self-assessment of the WDCC along the FAIR principles (Peters, Höck & Thiemann, 2020)[4] indicated a high level of FAIRness (0.9 of 1). That evaluation was purely based on self-developed metrics along the individual FAIR principles, did not evaluate individual datasets and provides a holistic view of the WDCC (meta)data curation approach.

Our study is further motivated by the fact that while it is clear that automation of FAIR-ness evaluation is needed for to ensure scaleability, we are unsure if automated tools are entirely fit-for-purpose, especially when it comes to the evaluation of contextual reusability of archived (meta)data (Wu et al., 2019; Bugbee et al., 2021; Dunn et al., 2021; Ganske et al., 2021; Murphy et al., 2021) - probably one of the most important aspects of "R". Or in other words: what use are good findability, accessibility and interoperability if the data lack contextual metadata like documentation of methods, uncertainty assessment, associated references or provenance information. We presume that automated assessment of such information is close to impossible with current technology - a question we address in detail in this study.

The aspect of contextual reusability is especially important to adequately consider when

[2]https://cera-www.dkrz.de/WDCC/ui/cerasearch/

[3]https://www.dkrz.de/en

[4]https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=fairness

assessing FAIRness of archived climate simulation data, because the climate modeling community has at least for the last decade provided access to standardized collections of well-documented data for reuse by the global community (Meehl et al., 2007; Taylor et al., 2012; Stockhause et al., 2012; Cinquini et al., 2014; Eyring et al., 2016; Stockhause & Lautenschlager, 2017; Balaji et al., 2018; Petrie et al., 2021). Since such efforts are only feasible by adhering to agreed upon and adopted discipline-specific (meta)data standards (e.g. Eaton et al., 2003; Ganske et al., 2021), this can already be seen as a certain degree of FAIRness. Further, data curation approaches of repositories catering for the archival of climate data already include quality control mechanisms to ensure long-term reusability (e.g., Stockhause et al. (2012), Evans et al. (2017) and Höck, Toussaint & Thiemann (2020)). FAIRness evaluation tools should therefore be capable of reflecting these efforts. In applying an ensemble of FAIRness evaluation tools in this study, we aim at answering the following research questions:

1. How does the previous self-assessment of WDCC FAIRness (Peters, Höck & Thiemann, 2020) compare to currently available tools and proposed methods, how is this reflected in WDCC's (meta)data curation approach and how can WDCC FAIRness be improved?

2. How do the different FAIRness evaluation tools compare to each other and what can we take home from such an analysis?

3. How fit-for-purpose are the different FAIRness evaluation tools for an evaluation of the domain-specific aspects of FAIRness, especially in terms of contextual (meta)data reusability?

Building on our analysis, we discuss the lessons-learned during the process of evaluation and conclude with a set of recommendations for the design and application of future FAIR evaluation approaches. The paper is organised as follows: we introduce our analysis method and data used in Section 2. This includes a detailed description of the FAIRness evaluation tools, the choice of evaluated WDCC-archived datasets and the approach taken to achieve comparability between the different FAIRness evaluation tools. Results are presented in Section 3 and discussed in Section 4. The paper concludes with a summary in Section 5.

2 Methods and Data

In this section, we detail our approach to selecting FAIRness evaluation tools for our ensemble from the pool of globally available tools. We also cover aspects of tool applica-

bility and discuss our approach to making the results from different tools comparable to each other. We also highlight the importance of constructive feedback-loops between tool developers and FAIRness evaluators. We further discuss and motivate our methodology behind the selection of WDCC-archived entries to be tested.

2.1 Selection of evaluation approaches

We based our selection of tools on the collection of FAIRness evaluation tools prepared by the Research Data Alliance (RDA) FAIR Data Maturity Working Group (WG)[5] (Bahim, Dekkers & Wyns, 2019). That collection presents twelve FAIR assessment tools having their origins at various institutions around the globe. We find that only two out of the twelve presented tools are actually fit-for-purpose in the context of our study. These are the *Checklist for Evaluation of Dataset Fitness for Use* (Austin et al., 2019) produced by the *Assessment of Data Fitness for Use WG (WDS/RDA)*[6] (cf. Sec. 2.1.1) and the *FAIR Maturity evaluation service* documented in Wilkinson et al. (2019) (cf. Sec. 2.1.2). The latter is not explicitly listed in Bahim, Dekkers & Wyns (2019), but represents the evolution of a listed tool (Wilkinson et al., 2018b). The other ten tools listed could either not be accessed (*ANDS-NECTAR RDS FAIR data assessment tool* and the *CSIRO 5-star Data Rating tool*), are not recommended to be used anymore by the creators (*DANS-Fairdat, DANS-Fair enough?* (L. Cepinskas, pers. comm. 24 March 21), did not provide clear and easy-to-use instructions regarding the tools' application (Peng et al. (2015); Peng et al. (2020); The MM-Serv Working Group (2018) and David et al. (2018)) or where from our perspective not suited for evaluation of FAIRness of a repositories' data holdings (Pergl et al., 2019).

We further sourced the internet by searching for "FAIR data evaluation". Thereby, we discovered the tool *FAIRshake* (Clarke et al., 2019) and decided to use it in our ensemble approach (cf. Sec. 2.1.3). We also discovered the ARDC's *FAIR self assessment tool* (Schweitzer et al., 2021), but decided not to use it as it neither provides a download option for test results annotated with sufficient metadata of the evaluated resource nor does it provide a quantitative measure of FAIRness as final output.

Building upon earlier collaboration with the developers of the F-UJI tool (Devaraju & Huber, 2020) (see examples in Devaraju et al., 2021), we also used that tool in its software version v1.1.1 for our assessment ensemble (cf. Sec. 2.1.4). Finally, we build on earlier in-house work to evaluate WDCC's FAIRness (Peters, Höck & Thiemann, 2020) and by performing a self-assessment using the metric collection presented in Bahim et al. (2020) (cf. Sec. 2.1.5).

[5] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[6] https://www.rd-alliance.org/groups/assessment-data-fitness-use

Tool	Acronym	method	Covered FAIR dimensions	Reference
Checklist for Evaluation of Dataset Fitness for Use	CFU	manual	n/a	Austin et al. (2019)
FAIR Maturity Evaluation Service	FMES	automated	F: 8, A: 5, I: 7, R: 2	Wilkinson et al. (2019)
FAIRshake	n/a	hybrid	F: 3, A: 1, I: 0, R: 5	Clarke et al. (2019)
F-UJI	n/a	automated	F: 7, A: 3, I: 4, R: 10	Devaraju et al. (2021)
Self Assessment	n/a	manual	F: 13, A: 12, I: 10, R: 10	Bahim et al. (2020)

Table 1: Summary of the five FAIRness evaluation tools used in this study. The hybrid method of FAIRshake combines automated and manual evaluation. The covered FAIR ((F)indable, (A)ccessible, (I)nteroperable, (R)eusable) dimensions refer to the number of metrics the tool tests, e.g. FMES checks for Findability using 8 different tests.

We summarize the main characteristics of the five FAIRness evaluation tools in Table 1. The detailed results obtained from applying the FAIRness evaluation approaches are available as supporting data (Peters-von Gehlen, 2021; Peters-von Gehlen & Hoeck, 2021).

2.1.1 Checklist for Evaluation of Dataset Fitness for Use (Austin et al., 2019)

The Checklist for Evaluation of Dataset Fitness for Use (CFU) was originally developed to supplement the CoreTrustSeal repository certification process (Austin et al., 2019) by providing a tool to "...check the fitness for use (e.g. FAIRness) of a repository's holdings..." (J. Petters, pers. comm., April 2021). So although not specifically designed with the FAIR principles in mind, CFU can be used in the context of our study because it addresses data curation aspects relevant in the context of FAIR.

The CFU is a manual questionnaire provided in the format of a google-form and can be accessed from the URL provided in Austin et al. (2019). The questionnaire consists of twenty questions covering aspects of dataset identification, state of the repository's certification, data curation, metadata completeness, accessibility, data completeness and correctness as well as findability and interoperability. It is evident, that the topics covered by the questions map very well onto the FAIR principles (Wilkinson et al., 2016). The questions allow for nuanced answers (Yes; Somewhat; No) and are formulated in a sufficiently generic way to allow for discipline-specific answers. Like for any manual questionnaire, the evaluator has to be familiar with the common practice of the scientific domain and, ideally, be aware of the repositories' preservation practice. The answers are saved to an online spreadsheet. Evaluators using the CFU can always come back to previous assessments, given that the spreadsheet is available, and comprehend the score a particular resource has attained. Objectiveness of an evaluator is key for reproducibility,

though. The provision of resource metadata in the form facilitates the findability and the results of an assessment can be shared with anyone.

2.1.2 FAIR Maturity Evaluation Service (Wilkinson et al., 2019)

The FAIR Maturity Evaluation Service (FMES) is a fully-automated FAIRness evaluation tool building on community-driven efforts in compiling discipline specific FAIR maturity indicators (Wilkinson et al., 2018a; Wilkinson et al., 2019). The current implementation of the FMES is accessible online[7] and lets users choose from a set of different FAIR maturity indicator collections for testing. At the time of writing, the majority of available collections is discipline agnostic and is provided by the tool developers.

For testing, the FMES takes the URL or PID of the online resource as input for finding and accessing the resource via the machine-actionable metadata profided as JSON-LD. If available, the PID strictly has to be provided to FMES to yield meaningful evaluation results (M. Wilkinson, pers. comm., April 2021). For later identification of the test, FMES also requires a title for the evaluation and the ORCiD of the evaluator as metadata. Once an evaluation has been performed - this can take up to 15 minutes to complete, we experienced an average of about 2 minutes per entry - the result of the evaluation is immediately displayed in the web interface and reasons for failing certain tests are documented (see Wilkinson et al., 2019, for more information). Evaluation scores are given in number of passed, n, versus number of total tests.

Every evaluation performed with the FMES is saved in its backend and can be searched for and accessed at any later time by anyone via the web-GUI. This enables comprehensibility and reproducibility of the evaluation results.

Here, we applied the FMES using the collection *All Maturity Indicator Tests as of May 8, 2019*[8]. We used that collection because it contains tests for all aspects of the FAIR principles (cf. Table 1), was compiled by the maintainer of the tool and because no climate science specific FAIR maturity indicator collection was available at the time of testing.

2.1.3 FAIRshake (Clarke et al., 2019)

The FAIRshake tool takes a hybrid (combination of manual and automated) approach to assessing the FAIRness of digital resources (Clarke et al., 2019). FAIRshake can be accessed online[9] and was initially designed for use in biology-related disciplines. The

[7] https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/

[8] https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/collections/6

[9] https://fairshake.cloud

framework is intentionally kept generic enough to also be applicable to other disciplines (D. Clarke, pers. comm., April 2021). Like with FMES, FAIRshake can be used with a number of different FAIR metrics collections, the so-called *rubrics*, which differ in the amount of included FAIR metrics, in the type of resource to be evaluated or in the scientific discipline the rubric can be applied to.

Applying FAIRshake is open to anybody upon online registration. Once registered, users organize their evaluations in projects, which contain the results from the digital resource assessments. The assessment itself is done by providing the URL to the digital resource, as well as further metadata like title, description and type of resource for later reference. The automated part of the evaluation sources the machine-actionable JSON-LD metadata of the resource. For our assessments, we used the *FAIRshake dataset rubric*[10] because it contains the in our view most adequate set of FAIR metrics for the purpose of our study (cf. Table 1) and the most comprehensible test formulations.
In the *FAIRshake dataset rubric*, an automated approach is taken to evaluate the metrics relating to accessing the dataset landing page, accessing the data, contacts and licensing. The other metrics focusing on documentation of the data and its provenance, the repository the data is hosted in, versioning and citation of the dataset have to be answered manually. If an automated test fails because the required criteria encoded in the tool are not met, the test can still be amended manually. The results are given as nuanced answers (Yes (100% score); Yes, but (75%); No, but (25%); No (0%)). An evaluator can add additional information like URLs or free-text to justify the provided answer, which often requires the evaluator being familiar with the common practice of the scientific domain and also of the repositories' preservation practice. Through the combination of automated and manual metric assessment, FAIRshake offers the unique possibility of testing for generic aspects of the FAIR principles, while also catering for domain-specific requirements.

Every assessment performed with FAIRshake can be accessed by anybody from the tools' homepage, allowing for transparency and reproducibility. Our results are organized in the FAIRshake project *WDCC for DSJ*[11].

[10] https://fairshake.cloud/rubric/8/
[11] https://fairshake.cloud/project/132/

2.1.4 F-UJI (Devaraju & Huber, 2020)

F-UJI is an automated tool for the assessment of the FAIRness of research data developed in the framework of the FAIRsFAIR[12] project. Within the project, as set of metrics which follow the core FAIR principles was developed for use with F-UJI (Devaraju et al., 2020). F-UJI not only enquires the machine-actionable (meta)data available as JSON-LD via the research data object's landing page (specified by either URL or PID), but also harvests any available information on the hosting repository or the dataset itself from external resources. These external resources include established services like re3data[13], DataCite[14], the RDA Metadata Standards Catalog[15] or Linked Open Vocabularies[16]. This approach supports the automated evaluation of domain-specific FAIRness by leveraging the advantages of domain-specific over general repositories. For a more detailed description of F-UJI features, please refer to Devaraju & Huber (2020) and Devaraju et al. (2021).

F-UJI is free to be used by anyone and can be either installed locally (Devaraju & Huber, 2020) or applied using an online demo version[17]. The software behind the online demo corresponds to the most recent software version available for local installation (R. Huber, pers. comm., April 2021). Here, we take the most economic approach for applying F-UJI and relied on the assessments of the online demo version. F-UJI takes the URL to the landing page of the resource to be tested as only input. An assessment itself happens on the order of a few seconds and the results are displayed in a dashboard-like manner. The overall FAIRness score is given in %, with each of the metrics having equal weights in the calculation.

An evaluator can easily enquire the reasons behind passed or failed tests by clicking on the corresponding icons. The results of an assessment can however not be saved online, making the comprehension of an earlier assessment result only possible by re-executing the assessment. Of course, this only makes sense if the F-UJI software stack hasn't been updated in the meantime - which may indeed happen since F-UJI is still in development and constantly updated (see Sec. 2.1.6). We saved a PDF version of F-UJI's output to our local infrastructure and have made them availabe via the WDCC (Peters-von Gehlen, 2021). For a more systematic application of F-UJI, a local installation is more beneficial.

[12] https://www.fairsfair.eu
[13] https://www.re3data.org
[14] https://datacite.org
[15] https://rdamsc.bath.ac.uk
[16] https://lov.linkeddata.es/dataset/lov/
[17] https://www.f-uji.net

2.1.5 WDCC-developed self assessment along Bahim et al. (2020)

We constructed our own manual FAIRness evaluation tool by building on earlier in-house efforts to evaluate the FAIRness of the WDCC (Peters, Höck & Thiemann, 2020)[18] and the FAIR metrics recommended in (Bahim et al., 2020). By relying on third-party recommendations on FAIR metrics (Bahim et al., 2020), the present approach reduces the risk of leaving the evaluation open for individual interpretation - a major problem of manual FAIRness assessments (e.g. Mons et al., 2017; Jacobsen et al., 2020). Almost all of the maturity indicators listed in Bahim et al. (2020) were evaluated, regardless of them being classified as Essential, Important or Useful, in order to obtain the most complete FAIRness assessment possible (cf. Supplement). We also allow for nuanced answers per maturity indicator where this makes sense, i.e. while some indicators can only fail (0%) or pass (100%), others can attain values in the range of 0% to 100%. For the final score per evaluated WDCC-entry, every FAIR maturity indicator is given equal weights.

Like for any manual FAIRness evaluation tool (cf. Secs. 2.1.1 and 2.1.3), trustworthy and useful conduction of the evaluation requires a strong background in discipline-specific practices and standards, while also allowing for a high degree of domain-specificity. The evaluation results are saved in a spreadsheet on local hardware and made publicly available in conjunction with this publication.

2.1.6 The benefit of contacting the tool authors

In the process of conducting the FAIRness assessments for this study, we inevitably came in contact with the developers to enquire upon usability of the tool for our purposes (CFU, FAIRshake), unexpected results (FMES, F-UJI) or to recommend enhancements to the user experience (FAIRshake). Especially for FMES and F-UJI, quick turnaround times in email communication resolved issues very efficiently. In both cases, our enquiries have lead to improvements of the software by revealing bugs in the code or making the evaluation approaches more flexible, e.g. making the recognition of PIDs in the JSON-LD metadata case insensitive (FMES, M. Wilkinson, pers. comm., April 2021). An example from F-UJI would be that the tool now correctly identifies the resource type from information given in the JSON-LD metadata - which leads to one more test passed (R. Huber, pers. comm., April 2021).
For FAIRshake, we used the tools' GitHub page[19] to raise issues recommending improvements to the look and feel of the tool as well as the automated test routines. These recommendations were promptly adopted (usually within less than a working day).

[18] https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=fairness
[19] https://github.com/MaayanLab/FAIRshake/issues

2.2 Selection of WDCC entries for evaluation

The WDCC is a domain-specific long-term archiving service focusing on ensuring the long-term reusability of datasets relevant for simulation-based climate science. Therefore, the main focus lies on the preservation of datasets stemming from numerical simulations of Earth's climate. Additionally, datasets originating from observations, e.g. satellite data products, aircraft observations and in-situ measurements, are also preserved in WDCC but make up a relatively small fraction of the total data volume. Datasets preserved in the WDCC are required to comply with domain-specific (meta)data standards and file formats and be accompanied by rich and scientifically relevant metadata so as to ensure long-term reusability.

The total volume of datasets preserved in WDCC amounts to ≈3.1 PetaBytes (August 2021)[20]. The largest part is represented by climate model output stemming from globally coordindated model intercomparison efforts like the the global Coupled Model Intercomparison Project 5 (CMIP5, Taylor, Stouffer & Meehl, 2012) or regionalisations thereof produced within the Coordinated Regional Climate Downscaling Experiment (CORDEX, Giorgi, Jones & Asrar, 2009). Those datasets are highly standardised, because global intercomparison studies rely on the efficient reusability of produced data across user communities. Indeed, data reuse is high for these datasets, therefore justifying the standardisation effort (Pronk, 2019). Smaller datasets archived in WDCC are comprised of climate modeling or observational projects organised at project or institutional levels (e.g. Heinzeller et al. (2017); Jungclaus & Esch (2009) or Seifert (2020)) and research output forming the basis of academic publications (e.g. Klepp et al. (2017) or Mülmenstädt et al. (2018)).

The degree of data maturity (cf. Höck, Toussaint & Thiemann, 2020, for maturity criteria) required for archival in WDCC depends on whether or not a DOI is to be assigned to the archived data: data have to fulfill higher technical and scientific quality requirements if a DOI is to be assigned in the archival process (cf. Peters, Höck & Thiemann, 2020, and references therein).

Individual WDCC-archived datasets, i.e. files, are stored as parts of larger data collections - an approach broadly adopted in simulation-based climate science community (e.g. Evans et al., 2017) and which builds on the OAIS (Open Archival Information System, CCSDS (2012)) framework. In an OAIS, the archived information is organised in Archival Information Packages (AIPs), with two specialized AIP-types being the Archival Information Unit (AIU) and the Archival Information Collection (AIC). Broadly speaking, AICs de-

[20] https://cera-www.dkrz.de/WDCC/ui/cerasearch/statistics?type=database_size

Project acronym	Data summary	Project volume [TB]	DOI assigned	Creation date	Comments
IPCC-AR5_CMIP5	Coupled Climate Model Output, prepared following CMIP5 guidelines and basis of the IPCC 5th Assessment Report (2 AICs evaluated)	1655	yes and no	2012-05-31 and 2011-10-10	
CliSAP	Observational data products from satellite remote sensing (2 AICs evaluated)	163	yes and no	2015-09-15 and 2009-11-12	one collection with no data access
WASCAL	Dynamically downscaled climate data for West Africa	73	yes	2017-02-23	
CMIP6_RCM_forcing_MPI-ESM1-2	Coupled Climate Model output prepared as boundary conditions for regional climate models, prepared following CMIP6 experiment guidelines	51	yes	2020-02-27	
MILLENIUM_COSMOS	Coupled Climate Model of ensemble simlations covering the last millenium (800-2000AD)	47	no	2009-05-12	
IPCC_TAR_ECHAM4/OPYC	Coupled Climate Model Output, prepared to support the IPCCs 3rd Assessment Report	2.6	yes	2003-01-26	Experiment and dataset with DOI; First ever DOI assigned to data (Stendel et al., 2004)
Storm_Tide_1906_German_Bight	Numerical simulation of the 1906 storm tide in the German Bight	0.3	yes	2020-10-27	
COPS	Observational data obtained from radar remote sensing during the COPS (Convective and Orographically-Induced Precipitation Study) campaign	0.2	yes	2008-01-28	
HDCP2-OBS	Observations collected during the HDCP^2 (High Definition Clouds and Precipitation for Climate Prediction) project	0.06	yes	2018-09-18	
OceanRAIN	In-situ, along-track shipboard observations of routinely measured atmospheric and oceanic state parameters over global oceans	0.01	yes	2017-12-13 7	
CARIBIC	Observations of atmospheric parameters obtained from commercial aircraft equipped with an instrumentation container	7.7E-5	no	2002-04-27	

Table 2: WDCC projects selected for evaluation. The project acronyms can be directly used to search and find the evaluated projects using the WDCC GUI. The project volume in TB (third column) refers to the total volume of the entire project named in the first column. A full listing with more comprehensive information on the evaluated WDCC-entries is provided in the spreadsheets underlying this study (cf Supplement).

scribe a collection of AIUs which are combined in a meaningful way to enable discoverability. AIUs contain metadata describing the archived actual datasets, whereas AICs contain metadata describing the respective collection of AIUs.

In the WDCC, data collections are comprised of "entries", i.e. AIPs, which follow a strictly hierarchical structure[21]: the topmost level is the "project", followed by the the levels "experiment" (AIC), "dataset_group" (AIC) and "dataset" (AIU) (WDCC, 2016). Of these, the entry types project and dataset are mandatory, whereas the entry types experiment and dataset_group are used as organisational backbone of larger collections. At the WDCC, DOIs are assigned at the AIC-level only. This is done to i), keep reference lists in publications using WDCC-archived data clear and concise and ii), display the effort put into the creation of a data collection through a single citation with the aim to elevate the data publication to the level of a paper publication. However, some older data preserved in WDCC also have DOIs assigned at the AIU, i.e. the dataset, level (e.g. Stendel et al., 2005).

An evaluation of the entire WDCC-archive is evidently out-of-scope as it contains >1.3M datasets, with a total number of 1126 DOIs assigned at the time of writing (August 2021)[22]. We have therefore chosen to evaluate a sample of thirteen WDCC-archived AICs (see Table 2), resulting in a total of 32 evaluated AIPs (thirteen experiments, six dataset_groups, thirteen datasets). In the selection of the sample, we aimed at providing a representative assessment across the entire spectrum of WDCC-archived data collections covering various degrees of data maturity while at the same time providing a representative sample in terms of data volume. We evaluated two AICs for two projects (IPCC-AR5_CMIP5 and CliSAP) because data maturity is heterogeneous in these projects. One AIC was evaluated for the remaining nine chosen projects. The evaluation approach is detailed in the next section.

We consider the evaluated AICs (cf. Table 2) as representative for the data maturity level of the entire WDCC-project they are associated with, allowing us to extrapolate the results of our evaluation. Doing so, the cumulative data volume of the WDCC projects evaluated here amounts to ≈2PB (cf Tab. 2). The sample is representative of about 65% of WDCC-archived data. The remaining 35% are represented by a large number of smaller AICs for which testing would have been out-of-scope in the context of this study due to time constraints. The results obtained from the evaluation of our sample thus provide a good indication of overall WDCC-FAIRness. We note here, that some of the evaluated AICs were archived before the advent of the FAIR principles and therefore represent the long-established WDCC-approach to ensure long-term reusability of archived data collections.

[21]https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf

[22]https://cera-www.dkrz.de/WDCC/ui/cerasearch/statistics?type=database_doi

2.3 Evaluation approach

The granularity of data collections archived in the WDCC is motivated by providing the most appropriate level of data organisation for accessibility and reuse (see above). The amount and richness of metadata (contacts, references, parameter lists, quality assessment reports, free text summary, etc) differs starkly between the levels of granularity. Therefore, reporting the FAIRness of WDCC-archived data at the level of individual AIUs would not be informative. Hence, we provide results of our assessment at the AIC level, i.e. at the level of a WDCC data collection. Also, this is the only way to do justice to the domain-specific approach of organising climate science related simulation-based and observational datasets in larger collections (Evans et al., 2017; Ganske et al., 2020).
In practice, we assessed all AICs presented in Table 2 at the level of their AIUs and averaged the results at the AIC-level for all assessment approaches for reporting, but for our self-assessment (Sec. 2.1.5). For that approach, we performed the evaluation directly at the AIC-level.

2.4 Achieving comparability among evaluation approaches

The applied FAIRness evaluation tools all show a different number of maturity indicators, which are also differently distributed along the FAIR dimensions. In order to achieve comparability between the assessment approaches, we took a pragmatic approach and simply averaged the results over all maturity indicator tests per approach. We do so, because this approach is automatically applied for the two automatic assessment approaches (F-UJI and FMES). Where necessary, we normalized the results to yield a FAIR-score in the range between 0 and 1, indicating a low- or high-level of FAIRness, respectively.
We acknowledge the fact that this way of comparing the results of different FAIRness evaluation tools somewhat distorts the results, because the results per FAIR dimension are not equally weighted. However, we argue here that our study has the main focus of raising awareness for available FAIRness evaluation tools and highlighting the intricacies associated with applying them. In the end, the results of most tests compare well at the AIC-level (see next section).

3 Results

3.1 Mean scores of FAIR assessments

We show the calculated scores obtained from the five FAIRness evaluation tools along with some general statistics in Table 3. The calculated level of FAIRness strongly depends on the assessment method and the evaluated AIC. Overall, we obtain an ensemble mean FAIR score for the WDCC of 0.67, with individual results per applied FAIRness evaluation tool ranging from 0.5 to 0.88. The calculation of the mean FAIR score does not account for any weighting by data volume per AIC. Scores are mostly higher for the manual or hybrid approaches compared to the automated ones. This is mostly because the automatic FAIRness evaluation tools include checks on the actual data, which require the evaluated data to be openly accessible by the evaluation tool. Since almost all WDCC-archived data are open and free for use by anyone, but only accessible after authentication, the automatic tests requiring data access fail by design. The manual evaluation tools however allow for an evaluation of WDCC-archived datasets, since these can be accessed through *human intervention* (wording taken from Bahim et al., 2020). Metadata must be prepared accordingly for automated tools, e.g. in the JSON-LD, so that it can also be evaluated. We discuss further aspects behind the differences in FAIRness scores between the applied methods in Section 4.

At the AIC-level (column "∅ per project" in Table 3), the spread around the ensemble mean is slightly smaller, ranging from 0.43 to 0.76. AICs with DOI obtain the highest FAIR scores, with an AIC associated with the project CMIP6_RCM_forcing_MPI-ESM1-2, which has a DOI assigned and is comprised of data produced within the framework of the CMIP6 initiative (Eyring et al., 2016), scoring highest.

Consequently, AICs having no DOI assigned, such as MILLENIUM_COSMOS, score lower. The lowest score is determined for one of the CliSAP AICs (CliSAP, no DOI and no data accessible). While that AIC does provide ample metadata on the corresponding WDCC landing pages (see cf Supplement for details to find the tested AICs), the data is not accessible because the status of the AIC was never set to "completely archived" by WDCC staff. The lack of data accessibility can in this case only be pinpointed using the manual and hybrid approaches - the automatic ones fail to recognise this major shortcoming and therefore cannot be used to capture the actual data curation status. While such curation levels are rather the exception than the rule for the WDCC, we deliberately chose to include an AIC with no accessible data in our evaluation to analyse the entire spectrum of WDCC data curation levels and for checking whether the automated tools recognize this.

Summarising this part of our results, we find that all FAIRness evaluation tools can be

used to reliably distinguish between various degrees of (meta)data curation of AICs preserved in the WDCC and that for the most part, AICs preserved in the WDCC satisfy the majority of the FAIR maturity indicators addressed by the applied evaluation approaches.

3.2 Agreement between evalution approaches

Our ensemble approach to FAIRness evaluation also offers the unique opportunity to analyse the consistency between the assessment approaches at the AIC-level. To illustrate this, we computed the relative standard deviation, defined as the standard deviation of a sample divided by the mean of the sample ($\frac{\sigma}{\varnothing}$), at the AIC level (rightmost column of Table 3) and the cross-correlations between the tests at the WDCC-level shown in Table 4.
If the applied FAIRness evaluation tools show a small spread in determined FAIRness scores for a particular project, they show agreement and $\frac{\sigma}{\varnothing}$ is small. We find the lowest values for datasets having a DOI assigned and being associated with ample machine-readable relevant metadata, i.e. CMIP6_RCM_forcing_MPI-ESM1-2 (Steger et al., 2020) and Storm_Tide_1906_German_Bight (Meyer et al., 2021), or a dataset with a low-level of domain-specific maturity (CARIBIC). At the other end of the spectrum, the FAIRness evaluation tools disagree most for the CliSAP AIC for which no data is accessible - for the reasons we alluded to in the previous paragraph. We provide a more detailed discussion of the differences between test results in Section 4.

The cross-correlations between the applied FAIRness evaluation tools (Table 4) clearly indicate that the level of agreement strongly depends on the applied methodology (manual, hybrid or automated), irrespective of covered FAIR dimensions per approach (see Section 2.1). Generally, the results of manual or hybrid approaches compare better to each other than to the automated ones. Similarly, the two automated approaches (FMES and F-UJI) compare well. However, there is an exception: the results of our Self Assessment and the F-UJI tool also compare relatively well.
Summarising this part of our results, we find that at the AIC-level, the five evaluation approaches broadly agree on the level of FAIRness (with one notable exception, see above). At the WDCC-level, we find that the scores obtained from FAIRness evaluation tools taking an identical methodology (manual, hybrid or automated) also compare well to each other. Here, manual and hybrid approaches can be seen as applying the same evaluation methodology ("human expert knowledge") as compared to the purely automated tests.

Project acronym	**Self Assessment**	**CFU**	**FMES**	**F-UJI**	**FAIRshake**	$\varnothing$ per project	σ per project	$\frac{\sigma}{\varnothing}$ per project
IPCC-AR5_CMIP5	0.84	0.72	0.44	0.58	0.95	0.71	0.20	0.29
IPCC-AR5_CMIP5, no DOI	0.65	0.67	0.44	0.54	0.93	0.65	0.19	0.29
CliSAP	0.86	0.78	0.48	0.58	0.97	0.73	0.20	0.28
CliSAP, no data accessible	0.27	0.30	0.43	0.52	0.64	0.43	0.15	0.36
WASCAL	0.90	0.80	0.50	0.58	0.91	0.74	0.18	0.25
CMIP6_RCM_forcing_MPI-ESM1-2	0.86	0.85	0.57	0.62	0.92	0.76	0.16	0.21
MILLENNIUM_COSMOS	0.63	0.53	0.45	0.51	0.82	0.59	0.14	0.24
IPCC_TAR_ECHAM4/OPYC	0.82	0.63	0.50	0.64	0.89	0.70	0.16	0.23
Storm_Tide_1906_German_Bight	0.90	0.68	0.55	0.62	0.83	0.71	0.15	0.21
COPS	0.86	0.47	0.53	0.55	0.87	0.66	0.19	0.29
HDCP2-OBS	0.90	0.48	0.53	0.59	0.86	0.67	0.19	0.29
OceanRAIN	0.90	0.75	0.57	0.60	0.97	0.76	0.18	0.23
CARIBIC	0.62	0.70	0.50	0.54	0.82	0.64	0.13	0.20
$\varnothing$ (WDCC)	0.77	0.64	0.50	0.58	0.88	0.67	0.15	0.22

Table 3: Results of FAIR assessments of WDCC data holding using the ensemble of FAIRness evaluation tools detailed in Section 2.1. The scores per test are calculated as unweighted mean over all tested FAIR maturity indicators. The mean ($\varnothing$), standard deviation (σ) and relative standard deviation ($\frac{\sigma}{\varnothing}$) on a project basis (three rightmost columns) are calculated across the scores of the five FAIR assessment tools. The mean value representative for the WDCC ($\varnothing$ (WDCC), last row) is calculated for all values in the respective column of the table. See main text for more details. Results at finer granularity are provided in the supporting data (Peters-von Gehlen & Hoeck, 2021)

	Self Assessment	**CFU**	**FMES**	**F-UJI**	**FAIRshake**
Self Assessment	n/a	0.61	0.65	0.73	0.79
CFU		n/a	0.36	0.50	0.78
FMES			n/a	0.65	0.30
F-UJI				n/a	0.49
FAIRshake					n/a

Table 4: Cross-correlations between the scores per project obtained with the five FAIRness evaluation tools (Table 3).

4 Discussion

From the beginning, the FAIR data guiding principles have been defined as being first and foremost applicable to any research discipline (Wilkinson et al., 2016; Mons et al., 2017) and that it requires the effort of domain specialists to define FAIRness maturity indicators at a discipline-level (Wilkinson et al., 2019). Since consolidation processes on the definition of suitable indicators are still ongoing in the global RDM community, we have put as much focus on discipline-specific aspects in our evaluation of WDCC-preserved (meta)data as possible. Global data sharing and data reuse is an essential part of everyday climate science and the community has developed and adopted relatively sophisticated (meta)data standards to facilitate reuse (Meehl et al., 2007; Stockhause et al., 2012; Taylor et al., 2012; Eyring et al., 2016; Ganske et al., 2020, 2021). At WDCC, (meta)data is preserved with a focus on long-term reusability and is therefore required to adhere to these standards to a certain degree - we therefore anticipated a relatively high degree of FAIRness for preserved (meta)data.

In this section, we discuss the domain-specific aspects impacting our analysis of WDCC-FAIRness (Section 4.1) and the differences between and comparability of the different evaluation approaches (Section 4.2). Further, we present lessons learned (Section 4.3) and finish off with recommendations to inform the development and operationalisation of FAIRness evaluation (Section 4.4).

4.1 Data granularity

At WDCC, preserved data is organised in data collections following a strict top-down hierarchy (cf. Section 2.3), where each level in the hierarchy is identified by an entry ID and has its own landing page in the WDCC GUI. Initially, we planned to present results for each hierarchy level of an AIC (cf. Table 2), but realized soon in the process that this approach does not reflect the evaluation of domain-specific FAIRness in climate science in general and data curation practice at WDCC in particular. As outlined in Section 2.3, we did in fact test all AIUs of the AICs separately and then computed the average. Because the amount and content of machine-actionable metadata varies starkly between the AIC hierarchy-levels, especially the automated evaluation approaches yielded a range of FAIRness scores for the AIUs of a single AIC. For example, F-UJI computed a scores of 0.54 and 0.7 at the “dataset” and “experiment” levels, respectively, for CMIP6_RCM_forcing_MPI-ESM1-2. In this case, the DOI is assigned at the experiment level, automatically resulting in a higher score. However, both entities must not be considered separately, as on the one hand, the actual data is not available at the experiment

level. On the other hand, the dataset level lacks the contextual information required for reuse. These domain-specific particularities of data granularity can at the moment not be captured with automated FAIRness evaluation tools but should be considered if FAIRness evaluation and certification become mandatory (see Section 4.4).

4.2 Comparability of test results

The varying capacities of the different FAIRness evaluation tools became very apparent and transpired early in our analysis. While the automated approaches (FMES and F-UJI) are useful for the evaluation of the machine-actionable aspects of preserved (meta)data, they fail to capture the actual curation status of (meta)data preserved in WDCC. We shortly describe four examples illustrating this point:

- Datasets preserved in WDCC are accessible for free, but only after authentication. The machine actionable metadata (JSON-LD) contain an indicator regarding data accessibility ("isAccessibleForFree": true). While this is in full compliance with FAIR principle A1.2, the automated test yield failed tests. While this result is fully explainable (FMES and F-UJI check for dataset URLs which are deliberately not included in the JSON-LDs for security reasons), it does reveal a central shortcoming of the automated evaluation approaches and highlights the intricacies of exactly matching the syntax of machine-actionable content required to pass automated tests.

- In cases when data are actually not available, the information on the availability status of the data is only provided on the landing page and not as part of the machine-readable metadata. Therefore, the automated approaches evaluate these AICs exactly as the other tested WDCC-entries (data is not accessible, test failed), resulting in too high FAIRness scores.

- Contextual information is practically impossible to evaluate using automated approaches. As the main goal behind providing FAIR data is to foster their reuse, providing adequate references, documentation and provenance information is essential. The machine-readable qualifiers ("subjectOf") included in the JSON-LDs lead to associated publications or reports. Once such a reference is detected by an automated evaluation approach, the corresponding test is passed. However, the actual content of the linked reference cannot be checked - it could therefore be completely irrelevant in the context of the evaluated (meta)data. In the context of this study, the AIC HDCP2-OBS represents such a case.

- By virtue of their intended application, the automated evaluation approaches do not take any information provided on the human-readable landing pages into account. At the WDCC, these often contain ample information about the data, like dataset size and file format. These parameters are not included in the JSON-LD because schema.org-requirements are vaguely defined.

All of the above points pose no problem to manual or hybrid tools. However, including the "human factor" in the evaluation process may lead to inconsistencies. A further limitation of manual FAIRness evaluation tools is the obvious inability to check for machine-actionability. Since this is an essential component of FAIR data, checking just for the human-readable aspects of preserved (meta)data is just as impeding as only checking for the machine-actionable aspects. Or put in other words, automated FAIRness evaluation tools check for the technical FAIRness - or reusability - whereas manual approaches (can) check for the contextual/scientific reusability.

A further point worth discussing is the comparability of the different test results. As outlined in Section 2.1, the five FAIRness evaluation tools do not cover the four FAIR dimensions in a comparable manner: FMES puts little focus on R (2 of 22), FAIRshake is dominated by R (5 of 9), F-UJI is dominated by F and R (together 17 of 24) and our own self assessment following Bahim et al. (2020) puts equal emphasis on all FAIR dimensions and is far more comprehensive than the other approaches (45 tests, compared to 20, 22, 9 and 24 for CFU, FMES, FAIRshake and F-UJI, respectively). Since there exist no recommendations regarding the importance of individual FAIR dimensions - apart from F, which is seen as the single most important principle of the FAIR spectrum to enable data reuse (Mons et al., 2017) - and their weighting in an evaluation, we provide simple arithmetic means of the test results. Similar to the ensemble approach applied in simulation based climate science, where the ensemble mean over multiple models is usually a better representation of reality than the simulation of an individual model (Tebaldi & Knutti, 2007), we see an added-value in presenting the mean over all FAIRness evaluation tools as "WDCC-FAIRness" (Table 3) as compared to relying on just a single test. Of course, once FAIRness evaluation becomes standardised and an operational requirement for repositories and archives in order to be regarded as trusted in science, basing a certification on the results of an ensemble of tests is impractical. We therefore hope that the results we present here help the community converge towards standardised, broadly applicable and officially recommended FAIRness evaluation tools.

4.3 Lessons learned

The process of applying five different FAIRness evaluation tools has helped us judge the WDCC preservation practice, critically reflect on our internal workflow, indicate avenues for improving the FAIRness of our (meta)data holdings and develop a sound understanding for domain-specific FAIRness in climate science.

- Machine actionability of archived data need not be the priority for data collections in the climate sciences. The size of datasets archived at WDCC is often $\mathcal{O}(10^2)$TB and more. It is simply not practical to include URLs pointing to the actual datasets in the machine readable metadata, as this may incur both security and bandwidth issues. The WDCC is currently implementing a PID-system at the dataset level to increase Findability.

- Some of the automated tests could have been passed, if the information given in the machine-actionable metadata would have been as comprehensive as that supplied on the landing pages of archived datasets. One example would be the specification of the file format. At the moment, we do not provide this information in the JSON-LD, because in some cases, the actual file format is NetCDF, a standard open file format of the climate science community, but the files are packed as .zip or .tar archives for download. Note however, that these issues are rather minor and do not reduce the FAIRness of WDCC data holdings per se - including them would merely increase the FAIR score of the automated evaluation approaches.

- Archiving of climate science related data in data collections characterised by a strict top-down hierarchy which do not have PIDs assigned to every data file is a main characteristic of the discipline-specific standard procedure to make these data available to the community. Evaluating a collection in its entity is essential to fully characterise its FAIRness.

- Reaching out to the developers of the evaluation tools was essential to apply the tools correctly, comprehend the test results and even discover bugs in the tools' source code. Close communication and collaboration between the tool developers and those wishing to apply them can not be overrated and we wish to contribute further to their development and testing in the future.

- In the process of defining the sample of AICs to be tested, we discovered several ones in which the data is not available due to shortcomings in the WDCC archival workflow. We are at the moment sieving through the WDCC data holdings to find

and amend these AICs and make the data associated with them available to the community.

- Applying the manual evaluation approaches is far less straight forward compared to the automated ones. Even if domain and repository experts perform the evaluation, the results may differ because subjectivity cannot be ruled out. One example would be a maturity indicator demanding the provision of dataset and provenance documentation. While supplying links to a third-party online database containing this information would suffice for one evaluator, this might not be the case for another one. Therefore, evaluation results obtained by one evaluator should always be reviewed. In this context, the list of FAIR maturity indicators compiled by Bahim et al. (2020) helps to reduce the risk of unconscious bias because it provides very specific guidance for testing.

- For some AICs, documentation is provided in terms of README files or reports which are archived along with the data. However, these files are hard to find if a user is not familiar with the WDCC and does not know where to look. WDCC-efforts to improve the user experience in this regard are underway by providing more clear access to associated documents and by working towards a community-acceptance of the EASYDAB (EArth SYstem DAta Branding, Ganske et al., 2021) concept which allows users to clearly identify high-quality archived datasets.

4.4 Recommendations for future FAIRness evaluation tools

In the course of our analysis, it became apparent that none of the five applied FAIRness evaluation approaches was entirely fit-for-purpose to evaluate the WDCC data-holdings (cf. Section 4.2 and 4.3), but all of them have their individual strengths on which to build future FAIRness evaluation tools.

For future FAIRness evaluation tools, we recommend the development of capable hybrid approaches to capture both the technical and contextual reusability of preserved research data.

For the reasons we elaborated on above, automated FAIRness evaluation tools are very good at testing maturity indicators which allow for binary yes/no answers following a standardised protocol. Of the two approaches used here, F-UJI seems to be more mature and capable than FMES, but still fails to capture the actual curation status of WDCC data holdings. At that point, the manual part of a FAIRness evaluation would take over to reliably judge the contextual reusability of the preserved (meta)data. Our recommendation to include domain experts and to not only rely on automated approaches in the evaluation of FAIRness and general (meta)data quality is also in-line with recent work on the same

topic following a similar line of argument (Wu et al., 2019; Bugbee et al., 2021; Murphy et al., 2021).

In practice, we envision a hybrid approach similar to that of FAIRshake, but substantially more comprehensive. The tool would also include internal databases specifying domain-specific information, like standards, file formats or essential metadata fields specific to the discipline. In this context, the concepts of FMES and FAIRshake enabling the use of different sets of maturity indicator catalogs is very promising.

5 Summary

In this study, we have applied an ensemble of five different FAIRness evaluation tools to evaluate the FAIRness of (meta)data preserved in the WDCC (World Data Center for Climate). The tools differed in terms of their applied methodology (manual, hybrid or automated evaluation) as well as in the weighting of the individual FAIR dimensions (Findable, Accessible, Interoperable or Reusable) in the evaluation. The research questions of our study were three-fold. First, the results of an earlier self-assessment of WDCC-FAIRness (Peters, Höck & Thiemann, 2020)[23] were to be compared to results from available third-party FAIRness evaluation tools and methods, including a further development of our self assessment approach. Second, we performed a comparative analysis of the results provided by the five tools to identify common strengths and/or weaknesses. Third, we intended to analyse the fitness-for-use of available FAIRness evaluation tools for the purpose of performing a comprehensive assessment of a repositories' (meta)data holdings. Building on the results of our study, the ultimate goals were to determine how WDCC's preservation guidelines live up to external FAIRness evaluation, to identify possible limitations and shortcomings and to provide recommendations to the global research data management community regarding the further development and application of FAIRness evaluation tools.

Addressing the first research question, we found that our previous self-assessment (Peters, Höck & Thiemann, 2020)[24] yielded a significantly higher level of WDCC-FAIRness (0.9 of 1) compared to the ensemble mean score of 0.67, with a range of 0.5 to 0.88, obtained from the five evaluation approaches applied here. Specifically, our self-assessment of this study, conducted along the recommendations of Bahim et al. (2020), yielded a lower score (0.77) than the previous one. We attribute this difference to the more comprehensive and objective evaluation presented in this paper. The web resource detailing WDCC FAIRness will be updated accordingly.

[23] https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=fairness

[24] https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=fairness

Regarding the second research question, we found tools involving manual assessment yield higher FAIRness scores than automated tools. This is because the automated approaches cannot be used to assess the contextual reusability of preserved (meta)data. As data in WDCC is preserved with a focus on long-term reusability, data is usually accompanied by rich metadata providing, for example, documentation and provenance information (Höck, Toussaint & Thiemann, 2020; WDCC, 2016) - an aspect which can only be adequately evaluated in a manual manner by a domain and/or repository expert. Further, lower FAIRness scores obtained from automated tools result from inaccessible data (WDCC data is only accessible after login, but for free) or missing information in the machine-actionable metadata provided by the WDCC. We are in the process of increasing the information content of those metadata. Further, the applied evaluation tools compare well at the data collection level if similar evaluation methodologies (manual, hybrid or automated) are used. An exception to this rule is the particularly good agreement between results from the automated F-UJI tool (Devaraju et al., 2021) and our own self-assessment based on Bahim et al. (2020). At the data collection level, we confirmed that a high-level of (meta)data maturity (Höck, Toussaint & Thiemann, 2020) also directly translates into high FAIR scores (and vice versa) across all FAIRness evaluation tools.

Regarding the third research question, we concluded that none of the five applied FAIRness evaluation tools provides a completely satisfactory evaluation experience by itself, because manual and automated approaches lack the capacity to quantify the machine- and contextual reusability of archive data, respectively. The hybrid methodology applied in FAIRshake (Clarke et al., 2019) is most promising in this regard as it merges the two approaches, but it lacked comprehensiveness in the setup we applied here.

Finally, we recommend to focus the development, application and operationalisation of future FAIRness evaluations on hybrid methodologies featuring a capable and comprehensive automated part and a contextual part evaluated by a domain and/or repository expert. Our recommendation is in-line with that of other recent studies (Wu et al., 2019; Bugbee et al., 2021; Murphy et al., 2021). We further strongly recommend that any part of a FAIRness evaluation be subject to scrutiny by expert reviewers.

With the ever increasing demand for archives and repositories to showcase their FAIRness, we see our results and recommendations a step forward to effectively consolidate efforts to develop and provide the most fit-for-purpose tools to evaluate discipline-specific FAIRness of digital objects.

Reproducibility

The data and methods underlying this study are made publicly available via the WDCC (Peters-von Gehlen, 2021; Peters-von Gehlen & Hoeck, 2021) and can be used to comprehend and reproduce the resuts presented here.

References

Austin, C., Cousijn, H., Diepenbroek, M., Petters, J. and Soares E Silva, M., 2019, WDS/RDA Assessment of Data Fitness for Use WG Outputs and Recommendations, doi:10.15497/rda00034.

Bahim, C., Dekkers, M. and Wyns, B., 2019, Results of an Analysis of Existing FAIR assessment tools, doi:10.15497/RDA00035.

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K. and Stall, S., 2020, The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments, Data Sci. J., 19, 41, doi:10.5334/dsj-2020-041.

Balaji, V., Taylor, K. E., Juckes, M., Lawrence, B. N., Durack, P. J., Lautenschlager, M., Blanton, C., Cinquini, L., Denvil, S., Elkington, M., Guglielmo, F., Guilyardi, E., Hassell, D., Kharin, S., Kindermann, S., Nikonov, S., Radhakrishnan, A., Stockhause, M., Weigel, T. and Williams, D., 2018, Requirements for a global data infrastructure in support of CMIP6, Geosci. Model Dev., 11, 3659–3680, doi:10.5194/gmd-11-3659-2018.

Bugbee, K., le Roux, J., Sisco, A., Kaulfus, A., Staton, P., Woods, C., Dixon, V., Lynnes, C. and Ramachandran, R., 2021, Improving Discovery and Use of NASA's Earth Observation Data Through Metadata Quality Assessments, Data Science Journal, 20, 17, doi:10.5334/dsj-2021-017.

CCSDS: REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS), RECOMMENDED PRACTICE, CCSDS 650.0-M-2 (Magenta Book), Issue 2, CCSDS Secretariat, Space Communications and Navigation Office, 7L70 Space Operations Mission Directorate, NASA Headquarters, Washington, DC, 20546-0001, USA, available at `https://public.ccsds.org/Pubs/650x0m2.pdf`, accessed 2021-06-14, 2012.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S.

and Schweitzer, R., 2014, The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, Future Gener. Comp. Sy., 36, 400–417, doi: 10.1016/j.future.2013.07.002.

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., Schilder, B. M., Robasky, K., Castillo, C., Idaszak, R., Ahalt, S. C., Williams, J., Schurer, S., Cooper, D. J., de Miranda Azevedo, R., Klenk, J. A., Haendel, M. A., Nedzel, J., Avillach, P., Shimoyama, M. E., Harris, R. M., Gamble, M., Poten, R., Charbonneau, A. L., Larkin, J., Brown, C. T., Bonazzi, V. R., Dumontier, M. J., Sansone, S. A. and Ma'ayan, A., 2019, FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources, Cell systems, 9, 417–421, doi:10.1016/j.cels.2019.09.011.

David, R., Mabile, L., Yahia, M., Cambon-Thomsen, A., Archambeau, A.-S., Bezuidenhout, L., Bekaert, S., Bertier, G., Bravo, E., Carpenter, J., Cohen-Nabeiro, A., Delavaud, A., De Rosa, M., Dollé, L., Grattarola, F., Murphy, F., Pamerlon, S., Specht, A., Tassé, A.-M., Thomsen, M. and Zilioli, M., 2018: Comment opérationnaliser et évaluer la prise en compte du concept "FAIR" dans le partage des données: vers une grille simplifiée d'évaluation du respect des critères FAIR., doi: 10.5281/zenodo.1995646.

Devaraju, A. and Huber, R., 2020, F-UJI - An Automated FAIR Data Assessment Tool, doi:10.5281/zenodo.4063720.

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J. and White, A., 2020, FAIRsFAIR Data Object Assessment Metrics, doi:10.5281/zenodo.4081213.

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Akerman, V., L'Hours, H., Davidson, J. and Diepenbroek, M., 2021, From Conceptualization to Implementation: FAIR Assessment of Research Data Objects, Data Sci. J., 20, 4, doi: 10.5334/dsj-2021-004.

Dillo, I. and de Leeuw, L., 2018, CoreTrustSeal, Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare, 71, 162–170, doi: 10.31263/voebm.v71i1.1981.

Dunn, R., Lief, C., Peng, G., Wright, W., Baddour, O., Donat, M., Dubuisson, B., Legeais, J.-F., Siegmund, P., Silveira, R., Wang, X. L. and Ziese, M., 2021, Stewardship maturity assessment tools for modernization of climate data management, Data Sci. J., 20, 7, doi:10.5334/dsj-2021-007.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Horne, R., Whiteaker, T., Blodgett, D., Zender, C. and Lee, D., 2003: NetCDF Climate and Forecast (CF) metadata conventions.

Evans, B., Druken, K., Wang, J., Yang, R., Richards, C. and Wyborn, L., 2017, A Data Quality Strategy to Enable FAIR, Programmatic Access across Large, Diverse Data Collections for High Performance Data Analysis, Informatics, 4, 45, doi: 10.3390/informatics4040045.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. and Taylor, K. E., 2016, Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, doi:10.5194/gmd-9-1937-2016.

Ganske, A., Heydebreck, D., Höck, H., Kraft, A., Quaas, J. and Kaiser, A., 2020, A short guide to increase FAIRness of atmospheric model data, Meteorol. Z., 29, 483–491, doi:10.1127/metz/2020/1042.

Ganske, A., Kraft, A., Kaiser, A., Heydebreck, D., Lammert, A., Höck, H., Thiemann, H., Voss, V., Grawe, D., Leitl, B., Schlünzen, K. H., Kretzschmar, J. and Quaas, J., 2021: ATMODAT Standard (v3.0), doi:10.35095/WDCC/atmodat_standard_en_v3_0.

Genova, F., Aronsen, J. M., Beyan, O., Harrower, N., Holl, A., Hooft, R. W., Principe, P., Slavec, A. and Jones, S.: Recommendations on FAIR metrics for EOSC, Publications Office of the European Union, doi:10.2777/70791, 2021.

Giorgi, F., Jones, C. and Asrar, G. R., 2009, Addressing climate information needs at the regional level: the CORDEX framework, World Meteorological Organization (WMO) Bulletin, 58, 175.

Heinzeller, D., Dieng, D., Smiatek, G., Olusegun, C., Klein, C., Hamann, I. and Kunstmann, H., 2017: WASCAL WRF60km with MPI-ESM_MR r1i1p1 forcing from the CMIP5 historical experiment, doi:10.1594/WDCC/WRF60_MPIESM_HIST.

Höck, H., Toussaint, F. and Thiemann, H., 2020, Fitness for Use of Data Objects Described with Quality Maturity Matrix at Different Phases of Data Production, Data Sci. J., 19, 45, doi:10.5334/dsj-2020-045.

Jacobsen, A., de Miranda Azevedo, R., Juty, N. S., Batista, D., Coles, S. J., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T. A., Goble, C. A., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K. M., Heringa, J., Hooft, R. W. W., Imming,

M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L. O. B., Schneider, J., Strawn, G. O., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E. L., Wittenburg, P., Roos, M., Mons, B. and Schultes, E., 2020, FAIR principles : interpretations and implementation considerations, Data Intelligence, 2, 10–29, doi:10.1162/dint_r_00024.

Jungclaus, J. and Esch, M., 2009: mil0021: MPI-M Earth System Modelling Framework: millennium full forcing experiment using solar forcing of Bard, URL `http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=mil0021`.

Klepp, C., Michel, S., Protat, A., Burdanowitz, J., Albern, N., Louf, V., Bakan, S., Dahl, A. and Thiele, T., 2017: Ocean Rainfall And Ice-phase precipitation measurement Network - OceanRAIN-W, doi:10.1594/WDCC/OceanRAIN-W.

Kruk, J., 2013, Good scientific practice and ethical principles in scientific research and higher education, Central European Journal of Sport Sciences and Medicine, 1, 25–29.

L'Hours, H., von Stein, I., Huigen, F., Devaraju, A., Mokrane, M., Davidson, J., de Vries, J., Herterich, P., Cepinskas, L. and Huber, R., 2020: CoreTrustSeal plus FAIR Overview, doi:10.5281/zenodo.4003630.

Meehl, G., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J., Stouffer, R. and Taylor, K., 2007, The WCRP CMIP3 multi-model dataset: A new era in climate change research, B. Am. Meteorol. Soc., 88, 1383–1394.

Meyer, E., Scholz, R. and Tinz, B., 2021: Reconstruction of the 1906 Storm Tide in the German Bright using TRIM-NP, FES2004, and DWD weather data, doi: 10.26050/WDCC/storm_tide_1906_DWD_reconstruct.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B. and Wilkinson, M. D., 2017, Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud, Information services & use, 37, 49–56, doi:10.3233/ISU-170824.

Mülmenstädt, J., Sourdeval, O., Henderson, D. S., L'Ecuyer, T. S., Unglaub, C., Jungandreas, L., Böhm, C., Russell, L. M. and Quaas, J., 2018: Using CALIOP to estimate cloud-field base height and its uncertainty: the Cloud Base Altitude Spatial Extrapolator (CBASE) algorithm and dataset, doi:10.1594/WDCC/CBASE.

Murphy, F., Bar-Sinai, M. and Martone, M. E., 2021, A tool for assessing alignment of biomedical data repositories with open, FAIR, citation and trustworthy principles, PLOS ONE, 16, 1–22, doi:10.1371/journal.pone.0253538.

Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A. and Ansari, S., 2015, A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets, Data Sci. J., 13, 231–253, doi:10.2481/dsj.14-049.

Peng, G., Wright, W., Baddour, O., Lief, C. and Group, T. S.-C. W., 2020, The WMO Stewardship Maturity Matrix for Climate Data (SMM-CD), doi:10.6084/m9.figshare.7006028.v11.

Pergl, R., Hooft, R. W. W., Suchánek, M., Knaisl, V. and Slifka, J., 2019, "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning, Data Sci. J., 18, 59, doi:10.5334/dsj-2019-059.

Peters, K., Höck, H. and Thiemann, H., 2020, FAIR long term preservation of climate and Earth System Science data with focus on reusability at the World Data Center for Climate (WDCC), Earth and Space Science Open Archive, p. 13, doi:10.1002/essoar.10501879.1.

Peters-von Gehlen, K., 2021: F-UJI evaluation output for the paper "Recommendations for discipline-specific FAIRness evaluation derived from applying an ensemble of evaluation tools", doi:10.35095/WDCC/F-UJI_results_WDCC.

Peters-von Gehlen, K. and Hoeck, H., 2021: Data underlying the publication "Recommendations for discipline-specific FAIRness evaluation derived from applying an ensemble of evaluation tools", doi:10.35095/WDCC/Results_from_FAIRness_eval.

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., Antonio, F., Berger, K., Bretonnière, P.-A., Cinquini, L., Dart, E., Dwarakanath, P., Druken, K., Evans, B., Franchistéguy, L., Gardoll, S., Gerbier, E., Greenslade, M., Hassell, D., Iwi, A., Juckes, M., Kindermann, S., Lacinski, L., Mirto, M., Nasser, A. B., Nassisi, P., Nienhouse, E., Nikonov, S., Nuzzo, A., Richards, C., Ridzwan, S., Rixen, M., Serradell, K., Snow, K., Stephens, A., Stockhause, M., Vahlenkamp, H. and Wagner, R., 2021, Coordinating an operational data distribution network for CMIP6 data, Geosci. Model Dev., 14, 629–644, doi:10.5194/gmd-14-629-2021.

Pronk, T. E., 2019, The time efficiency gain in sharing and reuse of research data, Data Sci. J., 18, 10, doi:10.5334/dsj-2019-010.

Schweitzer, M., Levett, K., Russell, K., White, A. and Unsworth, K., 2021: au-research/FAIR-Data-Assessment-Tool: Release v1.0, doi:10.5281/zenodo.4971127.

Seifert, P., 2020: HD(CP)2 short term observation data of Cloudnet products, HOPE campaign by LACROS, doi:10.26050/WDCC/HOPE_LACROS_CLN.

Steger, C., Schupfner, M., Wieners, K.-H., Wachsmann, F., Bittner, M., Jungclaus, J., Früh, B., Pankatz, K., Giorgetta, M., Reick, C., Legutke, S., Esch, M., Gayler, V., Haak, H., de Vrese, P., Raddatz, T., Mauritsen, T., von Storch, J.-S., Behrens, J., Brovkin, V., Claussen, M., Crueger, T., Fast, I., Fiedler, S., Hagemann, S., Hohenegger, C., Jahns, T., Kloster, S., Kinne, S., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Müller, W., Nabel, J., Notz, D., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Rast, S., Schmidt, H., Schnur, R., Schulzweida, U., Six, K., Stevens, B., Voigt, A. and Roeckner, E., 2020: CMIP6 ScenarioMIP DWD MPI-ESM1-2-HR ssp585_r2i1p1f1 - RCM-forcing data, doi:10.26050/WDCC/RCM_CMIP6_SSP585-HR_r2i1p1f1.

Stendel, M., Schmith, T., Roeckner, E. and Cubasch, U., 2004: ECHAM4_OPYC_SRES_A2: 110 YEARS COUPLED A2 RUN 6H VALUES, doi:10.1594/WDCC/EH4_OPYC_SRES_A2.

Stendel, M., Schmith, T., Roeckner, E. and Cubasch, U., 2005: EH4_OPYC_SRES_A2_APRS, doi:10.1594/WDCC/EH4_OPYC_SRES_A2_APRS.

Stockhause, M. and Lautenschlager, M., 2017, CMIP6 data citation of evolving data, Data Science Journal, 16, doi:10.5334/dsj-2017-030.

Stockhause, M., Höck, H., Toussaint, F. and Lautenschlager, M., 2012, Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data, Geosci. Model Dev., 5, 1023–1032, doi:10.5194/gmd-5-1023-2012.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A., 2012, An overview of CMIP5 and the experiment design, B. Am. Meteorol. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C. and Knutti, R., 2007, The use of the multi-model ensemble in probabilistic climate projections, Philos. T. Roy. Soc. A, 365, 2053–2075, doi:10.1098/rsta.2007.2076.

The MM-Serv Working Group, 2018, MM-Serv_ESIP_2018sum_2̌r1_20180709.pdf, doi:10.6084/m9.figshare.6855020.v1.

WDCC, 2016: CERA2 Metadata Submission Guide, https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf, accessed: 2021-06-09.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B., 2016, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data, 3, 1–9, doi: 10.1038/sdata.2016.18.

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., da Silva Santos, L. O. B., Prieto, M., McQuilton, P., Gautier, J., Murphy, D., Crosas, M. and Schultes, E., 2018b, Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework, bioRxiv, doi:10.1101/418376.

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Santos, L. O. B. D. S. and Dumontier, M., 2018a, A design framework and exemplar metrics for FAIRness., Sci. Data, 5, 180 118, doi:10.1038/sdata.2018.118.

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., da Silva Santos, L. O. B., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M. and Schultes, E., 2019, Evaluating FAIR maturity through a scalable, automated, community-governed framework, Sci. Data, 6, 1–12, doi:https://doi.org/10.1038/s41597-019-0184-5.

Wu, M., Psomopoulos, F., Khalsa, S. J. and de Waard, A., 2019, Data discovery paradigms: User Requirements and Recommendations for Data Repositories, Data Sci. J., 18, 3, doi:http://doi.org/10.5334/dsj-2019-003.