Engaging with Researchers and Raising Awareness of FAIR and Open Science through the FAIR+ Implementation Survey Tool (FAIRIST)

COLLECTION:
DATA MANAGEMENT PLANNING ACROSS DISCIPLINES AND INFRASTRUCTURES

PRACTICE PAPER

CHRISTINE R. KIRKPATRICK
KEVIN COAKLEY
JULIANNE CHRISTOPHER
INÊS DUTRA

*Author affiliations can be found in the back matter of this article

ABSTRACT

Seven years after the seminal paper on FAIR was published, that introduced the concept of making research outputs Findable, Accessible, Interoperable, and Reusable, researchers still struggle to understand how to implement the principles. For many researchers, FAIR promises long-term benefits for near-term effort, requires skills not yet acquired, and is one more thing in a long list of unfunded mandates and onerous requirements for scientists. Even for those required to, or who are convinced that they must make time for FAIR research practices, their preference is for just-in-time advice properly sized to the scientific artifacts and process. Because of the generality of most FAIR implementation guidance, it is difficult for a researcher to adjust to the advice according to their situation. Technological advances, especially in the area of artificial intelligence (AI) and machine learning (ML), complicate FAIR adoption, as researchers and data stewards ponder how to make software, workflows, and models FAIR and reproducible. The FAIR+ Implementation Survey Tool (FAIRIST) mitigates the problem by integrating research requirements with research proposals in a systematic way. FAIRIST factors in new scholarly outputs, such as nanopublications and notebooks, and the various research artifacts related to AI research (data, models, workflows, and benchmarks). Researchers step through a self-serve survey process and receive a table ready for use in their data management plan (DMP) and/or work plan. while gaining awareness of the FAIR Principles and Open Science concepts. FAIRIST is a model that uses part of the proposal process as a way to do outreach, raise awareness of FAIR dimensions and considerations, while providing timely assistance for competitive proposals.

CORRESPONDING AUTHOR:
Christine R. Kirkpatrick
San Diego Supercomputer Center, University of California San Diego, La Jolla, CA USA; Department of Computer Science, Faculty of Sciences of University of Porto, Porto, Portugal
christine@sdsc.edu

KEYWORDS:
FAIR; metadata; DMP; survey; reproducibility

TO CITE THIS ARTICLE:
Kirkpatrick, CR, Coakley, K, Christopher, J and Dutra, I. 2023. Engaging with Researchers and Raising Awareness of FAIR and Open Science through the FAIR+ Implementation Survey Tool (FAIRIST). *Data Science Journal*, 22: 32, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2023-032

1. INTRODUCTION AND BACKGROUND

The FAIR Principles describe 15 aspirational dimensions of research data management (Wilkinson 2016). They provide a starting point for mapping out data stewardship practices needed for any given research project. However, there is no way to ensure all research objects adhere to FAIR, and FAIR is not all encompassing. For example, FAIR is silent on data quality and reproducibility. FAIR also does not comment on data sovereignty, such as is covered in the CARE Principles of Indigenous Data Governance (Carroll 2020). Consensus is lacking on whose responsibility it is to ensure the FAIRness of research objects, as well as the underlying issue of who is responsible for research data (de Lima 2022; Wallis 2011). Stakeholders range from individual researchers, to institutions, funders and publishers (Bloemers 2020; Nicholson 2023). In each case, the stakeholder group is looking for clear and practical advice—and is less interested in philosophizing about the need for research data management practices or complex and detailed arguments over which approach is better. Researchers must address these topics only as much as funders require it. The US National Institutes of Health (NIH) require a data management plan and beginning in 2023, increased the requirements to cover data sharing (National Institutes of Health). The US National Science Foundation (NSF) requires a data management plan to be included with proposal materials. Funders around the globe require the discussion of research data and FAIR in varying degrees. Countries and regions that lead the trends include the European Union's (EU) research funding calls and Australia. Were it not for these funder requirements, researchers would only take these steps on a voluntary basis. However, this requirement provides a key opportunity for outreach and awareness of the FAIR principles and how they relate to newer technologies during proposal preparation. The FAIR+ Implementation Survey Tool (FAIRIST) creates information that can be included in a proposal's data management plan or project description. Its contribution and value are as much in what FAIRIST produces, as well as the conversations and decisions its completion evokes. Even where support and services are not available to researchers from their institution, the mention of FAIR implementation possibilities can initiate important discussion.

This work is organized as follows. Definitions and terminology are introduced in Section 2. Related work is also presented in this section. In Section 3, the motivations and stakeholders of FAIRIST are discussed. Section 4 provides detail on FAIRIST's design, functionality, and the user feedback process. This work closes with a discussion and perspectives on future work.

2. LITERATURE REVIEW, DEFINITIONS, AND TERMINOLOGY

This work refers to concepts from information and computer science.

FAIR data is data that meets principles in four categories: Findability, Accessibility, Interoperability and Reusability. The ***FAIR Principles*** are the 15 principles that correspond to making research objects FAIR (Wilkinson 2016). The principles are not prescriptive and are not rules, but rather are touchstones for concepts that lead to research object, or data, usability. One of the first steps in the FAIR Principles is to make data 'Findable' (LibGuides; Wilkinson 2016). Data should be easily findable both by humans and computers (Jablonka 2022; Vita 2018). The automatic and reliable discovery of datasets and services depends on machine-readable persistent identifiers and metadata. Persistent identifiers are important because they unambiguously identify data and facilitate data citation. An example would be a Digital Object Identifier (DOI). The (meta)data should be retrievable by their identifier using a standardized and open communications protocol, with restrictions in place if necessary. Metadata should be available even when the data are no longer available. Data do not need to be all open; they can be restricted and still be FAIR. Open or not, data should be stored somewhere safe for the long-term. The data should be able to be integrated with other data, applications, and workflows. The format of the data should therefore be open and interpretable for various tools. The concept of interoperability applies both at the data and metadata level. Common formats and standards and controlled vocabularies should be used. Ultimately, FAIR aims to optimize the reuse of data. To do this, data should be well-documented, have a clear license to govern the terms of its reuse, and provenance information.

FAIR Digital Objects: Even though the original FAIR Principles publication called for the need to make all types of research artifacts FAIR, there has been an overemphasis on data, e.g., a chunk of information or a single data point. This work acknowledges the need to make all digital objects FAIR, including software, models, algorithms, and workflows. The term 'FAIR Digital Object', or FDO, describes a concept and associated, evolving guidelines for packaging metadata about

each chunk of information and the data together—as well as associating each component with its own unique identifier. The complete FDO is assigned a master identifier to the assembled package of data and metadata (De Smedt 2020; Schultes 2019). FAIRIST takes the approach that all research objects should be assigned identifiers. In doing so, FAIRIST aims to move towards recommendations that provide advice to create FDO-compliant research objects.

FAIR+: To provide a shorthand for FAIR and reproducibility, FAIR+ is used as a term in this work.

Open Science has been used by different stakeholders to focus on different aspects of openness, from technological architecture to the (public) accessibility of knowledge creation, measurement and the democratization of access (Fecher 2014). This work focuses on the qualities of the research processes that lead to openness, including transparency and reproducibility through technical and practical approaches. FAIRIST supports open science aims via recommendations for implementing the FAIR Principles that relate to findability and accessibility, as well as reproducibility.

Reproducibility: This work uses Gundersen's definition of reproducibility. It chiefly states that science should be able to be reproduced, not to the extent that the results are numerically identical, but so that the results support the same inferences drawn from the original research (Gundersen 2018). Although often mistaken as the "R" in FAIR, reproducibility is aided by the implementation of the FAIR principles, especially those that pertain to the openness of software, tools and libraries, the accessibility of data, etc. Both FAIR and reproducibility are continuums, more than a destination. Research reproducibility can be resource-intensive, therefore researchers should do as much as possible to document and provide a path for another researcher to retest their conclusions. However, it is understood that it is often not possible to recreate the exact same environment, or to provide the compute and storage resources needed for reproducibility work.

RELATED WORK

The FAIR principles and associated literature give very good recommendations and guidance on how to address data in research projects. But usually, researchers work with text files and other non-structured documentation. Some efforts have been made to transform these recommendations and guidance into a computational tool that can standardize the process of data 'FAIRification'. Other important advancements that relate to FAIRIST include tools for interviewing researchers about FAIR implementation, data management practices, and upcoming tools for publishing and reusing data management plans. Four tools in particular were surveyed to understand if they could be extended to include the approach conceived for FAIRIST. Argos and DMPTool are both good candidates for partnership, as discussed in *Future Work*. The FAIR Implementation Profile Wizard uses a complementary approach, and it was important to understand what could be leveraged or learned from the platform. The last platform examined, FAIR Connect, is relevant as a potential platform for publishing and sharing plans created using FAIRIST.

Argos: A joint effort between OpenAIRE and EUDAT, this platform provides a way to create and manage Data Management Plans (DMP) (Argos). Argos allows for manual entry or guidance with a wizard. Argos aims to document a research project and its outputs, mainly datasets. The Argos UI is streamlined and appears to employ modern UI/UX principles. It uses FAIR principles in how it collects data, utilizing APIs wherever possible so that researchers, institutions, and funders are not manually entered but connected to a unique identifier. The Dataset feature allows a researcher to document a dataset manually or prefilled from templates, customized for the needs of the funder. Argos allows for collaborative writing, and DMP templates can be added, updated, and modified. Argos provides DOI and DMP versioning via Zenodo and supports export in JSON format. Argos does require knowledge of data management concepts to complete the forms. Argos is a potential partner for integrating some of the questions from FAIRIST, although it would be a significant expansion in scope for the platform.

DMPTool: This tool provided by the California Digital Library (CDL) was created several years ago to assist researchers with the creation of their data management plans to accompany proposals (Praetzellis 2019). The survey is comprehensive and updated regularly. However, the primary text is entered by researchers into many text boxes. This can be daunting for researchers who are new to research data management and are not sure where to start. Furthermore, in practice, researchers may only use the DMPTool once and then reuse plans from project to project, with minimal adjustment to the plan for the new project's needs. The survey and output of FAIRIST could be appended to the DMPTool and combined for ease of use by researchers.

FAIR Implementation Profile (FIP) Wizard: Created by members of the GO FAIR International Office and the GO FAIR Foundation, the FIP wizard eases the creation of a FAIR Implementation Profile that can be read by machines (FIP Wizard; FIP Wizard Documentation). One answers questions in survey style, and the output is in the Resource Description Framework (RDF) format. The tool is also part of an exercise for communities to discuss (metadata) standards choices, such as those used by the WorldFAIR project (WorldFAIR). Participants across several domains, or 'petals' of the project, reported that having the discussion around metadata choices was as valuable as creating the FIP (Law 2022). The FIP Wizard employs some of the same techniques and design as for FAIRIST. However, it is concerned with aggregate information for a domain or subdomain of science, rather than individual projects.

FAIR Connect: This new initiative from the GO FAIR Foundation and iOS Press seeks to extend new tools for data stewards and researchers (FAIR Connect). It provides a way to publish FIPs and DMPs as nanopublications. It also allows for data stewards to comment or endorse submissions. Additionally, it provides a way for stewards to be recognized for their contributions via citations. FAIRIST outputs could be published in FAIR Connect as nanopublications and assigned persistent identifiers for citation and attribution.

3. MOTIVATION AND STAKEHOLDERS

Researchers desire practical advice on how to implement the FAIR principles, but are challenged by the steep learning curve and background needed to engage in data stewardship. Some may not even be aware of FAIR until they see it mentioned in a funding solicitation. A systematic tool can help by narrowing down topics based on research activities and outputs planned, rather than the approach of presenting everything and leaving it to the researcher to select relevant principles. Such a tool should be designed to only broach topics that apply directly to the researchers' planned work. Those in the humanities who only plan to produce data and disseminate findings on a website would not encounter more complex topics, such as where to share their machine learning (ML) models. Conversely, a computationally intensive project will find specific suggestions on where they might deposit ML artifacts and how to aid the reproducibility of their work by others in the future.

FAIRIST began as a templated response used to assist colleagues in crafting DMPs. In particular, researchers sought advice on how to implement FAIR, how to address FAIR when machine learning is employed, as well as what artifacts to make FAIR. The implementation advice distilled as many of the FAIR principles as possible into a table that a researcher could include in their DMP. Table 1 gives an example of the text created for an NSF proposal. This template was reused for other proposals, where the project name was replaced and the dimensions of FAIR added

Table 1 Template that inspired the creation of FAIRIST.

FAIR DIMENSION	
Findable	• Data will be assigned a PID <how?> and will be referenced on the <project website> • A catalog entry will be added to <FAIR Data Point or community/institutional catalog>. • Metadata and links to related ontologies will be available on the <project website>. • Where tags exist, schema.org descriptors will be utilized.
Accessible	• Available via <storage location>, that doesn't require specialized software to access. This includes both the raw data and curated or derived data. • The surrogate and other ML benchmarks will be deposited in <repository>. • Any APIs will be versioned and described, linked from the <project website>.
Interoperable	• Code stored on github and linked from the <project website> • Uses libraries from <project name> that utilize <standard or standard Python libraries, etc.>. • Uses standard references for <more here>. • Both input and output data are in <specify> format.
Reusable	• ML model and data will be deposited at <repository>. • Notebooks will demonstrate how to assemble model and sample training datasets. Each notebook product will be assigned a DOI using <specify DOI source>. • The <project> notebook interface is on <place shared, e.g., github>. • Provenance of the simulation creation will be available as part of the metadata. • A designation will be added to the website noting that all data as licensed under Creative Commons Attribution 4.0 International License.

or subtracted depending on the planned research. This capitalized on researchers' interest in learning more about FAIR implementation and research data management during the proposal process. However, proposal development is a very busy time and, most attention is given to the project description or plan, not the DMP. Knowing this, the advice given was created to be almost ready for inclusion in the DMP, and areas to update were clearly marked in brackets (< >).

After filling out these templates manually a few times, it became clear that the process could be streamlined through a self-service survey. Even though every research project is different and the topics can be complex, much of the human logic could be distilled into 'if/then' statements. For example, if the project means to produce notebooks, then the DMP should specify where notebooks will be shared, if they will be given a DOI, and if a notebook template will be used.

FAIRIST provides customized text for a researcher to include in a data management plan or proposal. Some form of DMP is required by many federal funders. The added benefits of planning data management at the outset of a research project are many: it makes it easier to audit, to check compliance with requirements, and to document the project which all benefit both researchers and funding agencies. Raising topics as part of creating a required document can also put a research project in good stead for complying with other domain-specific publication requirements later. For example, the Association for the Advancement of Artificial Intelligence (AAAI) hosts one of the most prestigious annual conferences for AI researchers. Papers submitted must also include a reproducibility checklist (AAAI Conference 2022). Many of the implementation solutions to the FAIR principles aid researchers in also being ready for the reproducibility checklist. For example, for papers submitted to AAAI that rely on data sets, this question must be answered, 'All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA).' Adherence to the FAIR principles relating to the clear statement of data usage licenses and the accessibility of data would prepare researchers to answer 'yes' to this question.

The stakeholders for a tool like FAIRIST include researchers from all domains and sectors, like academia and industry, although FAIRIST is tuned for research grant proposals. Additional stakeholders include anyone involved in the proposal process where a DMP is required or where the discussion of the FAIR principles is beneficial. This could include research support professionals, pre-award and project managers, and students or postdocs involved in proposal creation. This tool could also be used in synchronous and asynchronous trainings, such as the CODATA-RDA Schools of Research Data Science curriculum (CODATA-RDA-DataScienceSchools/Materials), a grantsmanship course (NSF HSI National STEM Resource Hub), or a data management plan training course hosted by a university library.

4. FAIRIST TECHNOLOGY & TESTING

FAIRIST surveys aspects of the project and then maps them to possible options and suggestions. Based on past experience and project requirements, all topics in FAIRIST are organized such that the logic is easily followed by the researcher in an attempt to save them time. When the form is complete, all information is automatically generated and embedded in the project proposal in an organized and structured way. The important qualities of this approach include: it makes a complex topic accessible; makes efficient use of researchers' time; and uses the time spent in the survey to lift awareness of the topic, its richness and dimensions. For example, if the project being described will produce Machine Learning (ML) models, then a follow up question is added asking, 'Where will the ML models be shared?', with several answers the user may or may not be aware of previously. An additional question asks, 'What are the reproducibility considerations you will undertake to document analysis that utilizes ML?' (Figure 1). By providing check box options rather than only a free form text box, the user can gain knowledge about the topic that doesn't rely on specific understanding of FAIR+ concepts. The example shown in Figure 1 distills ML implementation factors that can affect reproducibility to introduce the concept and ways to remediate variability.

The reproducibility consideration options are distilled from a much longer and complex Computer Science paper on sources of irreproducibility (Gundersen 2022). The source paper is linked in the FAIRIST survey question, in case the user wishes to read more about the topic before deciding or implementing the suggestions. The advice from the paper was adapted as

postcard-sized material (Figure 2) that could be used for outreach and awareness building of the concept and FAIRIST tool (Kirkpatrick 2022). This approach could be used to rapidly put other research data management scholarship into practice.

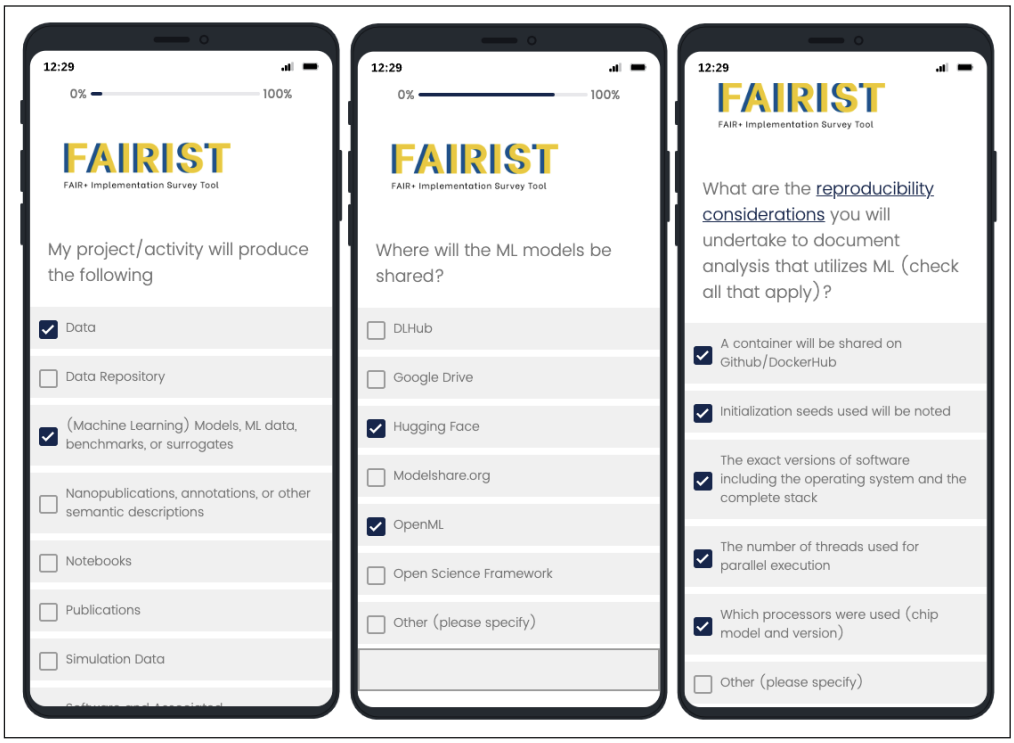

Figure 1 Embedded logic in FAIRIST expands the survey questions to fit the project described.

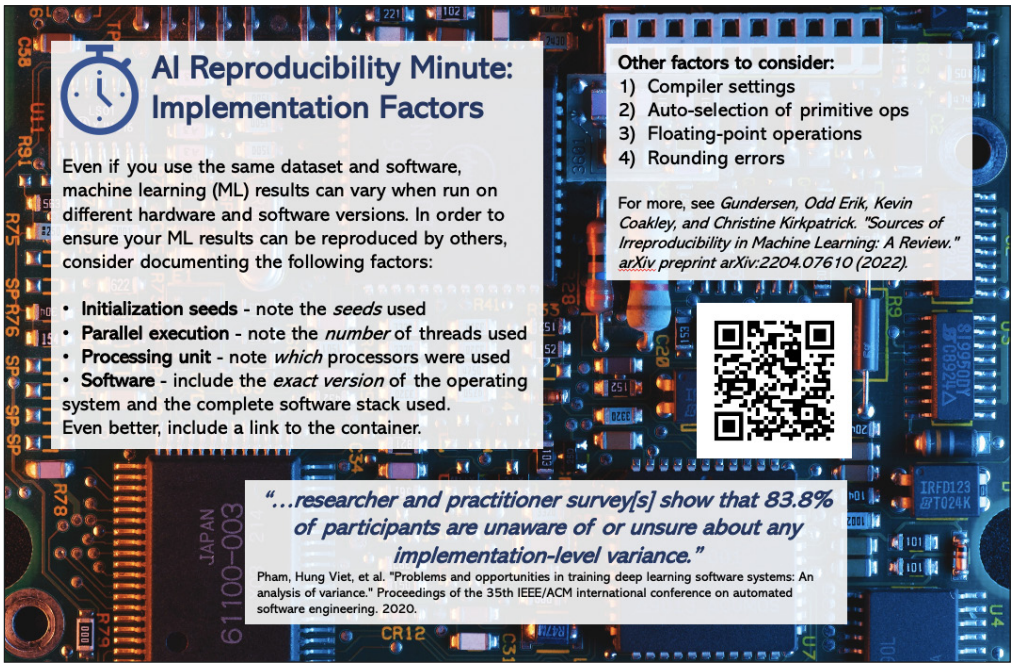

Figure 2 Outreach and awareness building material that could be used in concert with FAIRIST by libraries, research computing facilitators, and other researcher support personnel.

TECHNOLOGIES & METHODOLOGY UTILIZED

FAIRIST was built in Qualtrics, a business survey tool with a full-featured user interface (UI) that allows for survey customization as the input is given (Qualtrics). For example, it is possible to ask up front what types of research objects will be created and add or skip questions automatically based on the first response. The Qualtrics UI allows for a non-programmer to refine the form iteratively, enabling an 'Agile' approach. Agile refers to a software methodology with four pillars: 'individuals and interactions over tools and processes, working software over comprehensive documentation, customer collaboration over contract negotiation, and responding to change over following a plan' (Beck 2001). At the time Agile was introduced, it stood in stark contrast to process and resource-intensive methodologies, such as Waterfall

(McCormick 2012; Sureshchandra 2008). The key principles adopted in the creation of FAIRIST including focusing on the individual's needs above the assumed process of writing a DMP and creating a working example quickly that could be iteratively refined based on early user input.

Qualtrics is limited in the customized output it can provide. Some of the survey responses are used to determine what additional questions to ask, whereas the other responses set variables. Using the Qualtrics API, the variables call a Python script hosted on a local cloud instance that transforms the variables into completed sentences (Coakley 2022). This text output is formatted for inclusion in a DMP but could also be extended to include a machine actionable DMP. This would be accomplished by providing the text formatted for written language, as well as in RDF. The Python program that formats the variables collected from Qualtrics could be adapted to any format including the Research Data Alliance DMP common standard (Miksa 2020).

The Qualtrics web service workflow is asynchronous; it can take from 1–5 minutes for the output to be sent to the FAIRIST output generator. A user is notified by email when the FAIRIST output is ready to be retrieved. FAIRIST output links are encoded with a 128-bit universally unique identifier, so that others' output can't be easily guessed, and thus viewed. After a few minutes, an email is sent to notify the user that the recommendations are available for viewing. An example of the output is shown in Table 2.

Table 2 Example Output from FAIRIST Showing Recommendations.

FAIRIST Recommendations

Based on your responses, the following recommendations are included for your consideration and/or inclusion in your project's Data Management Plan.

Types of Data

Research objects associated with the project can be classified into the following groups:

- Data
- (Machine Learning) Models

Data Stewardship Practices Planned

Table 1 shows specific data stewardship actions that will be undertaken during the project as they relate to the high-level goals of FAIR.

FAIR DIMENSION	RESEARCH DATA STEWARDSHIP PRACTICES PLANNED
Findable	• Research products will be posted to the Project website. • Data will be assigned a unique identifier per community best practices and will be referenced on the Project's website. • Metadata and links to related ontologies will be available on the Project website. • Where tags exist, schema.org descriptors will be utilized.
Accessible	• Available via open, web accessible folder. • All data is open.
Interoperable	• Code stored on github (and linked from the Project website). • Uses libraries included with the code. • Both input and output data are in HDF5 format.
Reusable	• ML model and data will be deposited at OpenML.org. • A notice posted will designate research objects as licensed under CC-BY.

Table 1: Data Stewardship Practices Planned by FAIR Dimension

USER FEEDBACK AND TESTING

FAIRIST questions were developed within the core team and refined as a questionnaire in a document. Once that was converted to the Qualtrics format and initial user testing began, the wording of questions was refined for clarity and brevity. Several questions had to be reworded, so that the user input specified would form a grammatically correct sentence. For example, for the question, 'Will your data management plan, or the document this is being developed for, be shared?', the options were: 'Yes, at FAIR Connect'; 'Yes (specify)'; 'No'. In the feedback, the 'Yes' options triggered the inclusion of the sentence, 'This plan will be shared', appended by the variable for the question. The user input for 'Yes (specify)' would be a grammatically incorrect sentence, unless the user knew to begin with a preposition. If the user specified, 'my institutional repository', the resulting sentence would be 'The plan will be shared [sic]

my institutional repository'. This question was changed to not include a 'specify' option. If the feedback is 'Yes' the feedback includes, 'The plan will be shared.' The information on FAIR Connect was moved to be part of the question, 'Examples of places to share DMPs include FAIR Connect' with a link to the FAIR Connect website. Where possible, links were embedded into the questions, so that users could read more about a topic, for example, if a user answers 'Yes' to 'Will an API be provided?', they would receive the follow up question, 'Are you using a Smart API?', where 'Smart API' links to the website https://smart-api.info/.

Once initial tests were complete, the tool was released to researchers at University of California San Diego and the University of Porto. A webform was supplied for reporting bugs and feature requests. Several researchers sought out the project team and provided direct feedback. Most indicated that FAIRIST was easy to use and that they would recommend it to colleagues.

Highlights of the feedback:

- Some reported that FAIRIST made them aware of a new tool previously unknown, including Smart API mentioned above.
- One researcher wanted to 'reverse engineer' how to implement FAIR for his project, as his results only indicated he had plans for Findable and Accessible, but not Interoperable or Reusable.
- Another researcher noted that FAIRIST made no mention of gateways, a software portal used to access other infrastructure, especially to make High Performance Computing (HPC) more accessible.

One of the major takeaways was that FAIRIST's value rested in helping streamline a part of the proposal process. The major bugs uncovered related to the ability to specify custom text or "other" on most questions. This created extra complexity on the programming side and many of those responses were difficult to insert into fill-in-the-blank sentences. Based on the researchers' feedback:

- Some of the 'specify' answer options were or will be eliminated. Feedback that incorporates 'specify' responses will be rewritten to work with open-ended statements.
- Relevant questions now include an answer option for gateways.

FAIRIST is available for use at http://fairist.sdsc.edu/.

Feedback can be submitted at https://tinyurl.com/fairist.

5. CONCLUSION AND FUTURE WORK

While FAIR implementation is dependent on the specific factors for each research project and domain, as well as changes with evolving technology, it is possible to provide researchers with concrete advice. Tools such as FAIRIST provide a framework for embedding new information as practices develop and raise awareness on open science practices. By utilizing Agile methodologies and readily available cloud-based tools to create FAIR tools and resources, implementation advice can be more rapidly distilled and presented to researchers. These takeaways can be packaged not only for use in FAIRIST and tools like it, but also reformatted as outreach and awareness tools that promote both tools and FAIR+ concepts. Though created with the motivation to assist researchers and their teams with FAIR implementation, and to increase the adoption of the FAIR Principles, survey tools with proactive suggestions can assist researchers in other ways. Streamlined tools like FAIRIST can anticipate publishing and other funder requirements.

The preliminary results obtained from researchers are positive, and it seems FAIRIST is a good proof of concept. A more detailed evaluation is needed, where the refined FAIRIST obtained with the preliminary results is made available to a larger audience, along with the refined questionnaire. The analysis of this data will open new paths for improvement. A call for feedback will be issued through partnerships with research data consortia and other organizations active in both research and FAIR practice development. In the interim, feedback received from users will be considered and used to incrementally improve FAIRIST. The team is contemplating a feature that speaks to the request to reverse engineer the survey, which would allow one to view the full set of FAIR implementation steps. Further feedback from researchers will be gathered to inform what would be most useful and how it should be presented.

Future work should include a wider review of the survey options and outputs by experts in information science and research computing. It would be more sustainable if these questions and survey techniques were adopted by an existing tool. However, if that does not occur, an advisory board should be formed to guide decisions. For example, one of the questions, 'Where will your ML datasets be shared?' provides several options. Should the survey reflect the current practices or eliminate options the community determines to be suboptimal? Qualtrics as a platform is not a long-term solution for FAIRIST because of its output limitations: waiting for an email with a link to feedback rather than displaying the information immediately. However, in the short-term, Qualtrics enables rapid, incremental improvements based on user testing and input from other experts. Once FAIRIST's questions and output have been well tested and refined, FAIRIST should migrate to a custom, efficient, stand-alone Python script and/or be integrated with an existing DMP tool. If FAIRIST remains a stand-alone tool, it should be converted to an application with a database for storing surveys and accessing past FAIRIST output. The output should be immediately available and be also formatted as a machine-readable output using an established standard, in addition to the text meant for DMPs. Future work could include connecting FAIRIST to data sources, so that funder and program specifics could influence the questions asked and feedback given. It would also allow for the specification of resources by PID, such as specific equipment and standards. This is already present to some degree in Argo (*Related Work)*; a potential path would be to fold FAIRIST's features into an existing platform such as Argo.

There are numerous other sources to mine for potential questions and implementation suggestions. This work focused on Computer Science and AI because of the authors' backgrounds and a perceived gap in implementation advice. A future version of FAIRIST could include custom options tailored to advice for specific domains. This is beyond the complexity that can be handled by Qualtrics but would be possible in a future iteration of FAIRIST. For example, for projects that will produce a new domain repository and are from the Earth Sciences, FAIRIST could offer the option to include the repository in the Magnetics Information Consortium (MagIC) or the Council of Data Facilities (CDF) consortium (Council of Data Facilities 2022). At the moment, this option is offered to anyone, regardless of scientific domain, which indicates that the project will create a domain repository. FAIRIST could be further tailored to research needs by asking custom questions based on the agency funding source. As funders or institutions implement machine actionable DMPs, FAIRIST could also include the implementation guidance in machine readable format, e.g., triples in RDF. This could then be used to automatically verify compliance with the planned research data management practices.

ACKNOWLEDGEMENTS

Thanks to Melissa Cragin (Rice University) for her early feedback on the FAIRIST survey questions and to Odd Erik Gundersen (NTNU) for his mentorship regarding AI reproducibility. The FAIRIST logo was created by Alexandra Andrieu (SDSC, UC San Diego). Lynne Schreiber (SDSC, UC San Diego) co-created the outreach materials shown in Figure 2.

FUNDING INFORMATION

This work was partially funded by National Science Foundation (NSF) awards #2226453, 1928208, #1916481 and FAIR pilot funding support from the San Diego Supercomputer Center.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR INFORMATION

Christine R. Kirkpatrick is the division director of Research Data Services at the San Diego Supercomputer Center, UC San Diego. She heads GO FAIR US and is the Secretary General of CODATA.

Kevin Coakley is a Senior Cloud Integration Specialist at the San Diego Supercomputer Center, UC San Diego. Kevin is a PhD student in the computer science department at NTNU, specializing in AI reproducibility.

Julie Christopher is an IT project manager at the San Diego Supercomputer Center, UC San Diego.

Inês Dutra is an Assistant Professor in the Department of Computer Science, Faculty of Sciences of University of Porto, Portugal. Her main research interests are parallelism, logic programming and interpretable machine learning models.

AUTHOR CONTRIBUTIONS

C. Kirkpatrick wrote the practice paper, authored the survey tool, output, and co-created the outreach material in Figure 2. K. Coakley provided input on the paper, co-wrote the outreach material in Figure 2, and wrote the script and API interaction between Qualtrics and the cloud instance for custom output. J. Christopher formatted the survey in Qualtrics, co-created the outreach material in Figure 2, provided input on the paper, and finalized the references. I. Dutra provided advice on the project scope, added technical and structural input to the paper, wrote specific sections, and provided substantial feedback on FAIRIST and the paper.

AUTHOR AFFILIATIONS

Christine R. Kirkpatrick orcid.org/0000-0002-4451-8042
San Diego Supercomputer Center, University of California San Diego, La Jolla, CA USA; Department of Computer Science, Faculty of Sciences of University of Porto, Porto, Portugal

Kevin Coakley orcid.org/0000-0003-4976-8136
San Diego Supercomputer Center, University of California San Diego, La Jolla, CA USA; Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

Julianne Christopher orcid.org/0009-0006-4290-1810
San Diego Supercomputer Center, University of California San Diego, La Jolla, CA USA

Inês Dutra orcid.org/0000-0002-3578-7769
CINTESIS@RISE, Department of Computer Science, Faculty of Sciences of University of Porto, Porto, Portugal

REFERENCES

AAAI Conference. 2022. Reproducibility Checklist, 22 February–1 March 2022. Available at https://aaai.org/Conferences/AAAI-22/reproducibility-checklist/ [Last accessed 9 December 2022].

Argos. Plan and follow your data. Available at https://argos.openaire.eu/ [Last accessed 9 December 2022].

Beck, K, Beedle, M, Van Bennekum, A, Cockburn, A, Cunningham, W, Fowler, M, Grenning, J, Highsmith, J, Hunt, A, Jeffries, R and **Kern, J.** 2001. Manifesto for agile software development.

Bloemers, M and **Montesanti, A.** 2020. The FAIR funding model: providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices. *Data Intelligence*, 2(1–2): 171–180. DOI: https://doi.org/10.1162/dint_a_00039

Carroll, SR, et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal,* 19(1): 43. DOI: https://doi.org/10.5334/dsj-2020-043

Coakley, K. 2022. FAIR Matrix Survey API. Available at https://github.com/kevincoakley/fair-matrix-survey-api [Last accessed 15 December 2022].

CODATA-RDA-DataScienceSchools/Materials. Available at https://github.com/CODATA-RDA-DataScienceSchools/Materials/blob/master/docs/DataAtlanta2022/index.md [Last accessed 9 December 2022].

CODATA-RDA Schools of Research Data Science. Available at https://www.datascienceschools.org/ [Last accessed 9 December 2022].

Council of Data Facilities: Geoscience Research. Available at https://www.earthcube.org/council-of-data-facilities [Last accessed 9 December 2022].

De Lima, RA, Phillips, OL, Duque, A, Tello, JS, Davies, SJ, de Oliveira, AA, Muller, S, Honorio Coronado, EN , Vilanova, E, Cuni-Sanchez, A and **Baker, TR.** 2022. Making forest data fair and open. *Nature Ecology & Evolution*, 6(6): 656–658. DOI: https://doi.org/10.1038/s41559-022-01738-7

De Smedt, K, Koureas, D and **Wittenburg, P.** 2020. FAIR digital objects for science: From data pieces to actionable knowledge units. *Publications*, 8(2): 21. DOI: https://doi.org/10.3390/publications8020021

FAIR Connect. Available at https://fairconnect.pro/ [Last accessed 9 December 2022].

Fecher, B and **Friesike, S.** 2014. Open science: one term, five schools of thought. *Opening science*, 17–47. DOI: https://doi.org/10.1007/978-3-319-00026-8_2

FIP Wizard. Available at https://fip-wizard.ds-wizard.org/ [Last accessed 9 December 2022].

FIP Wizard Documentation. Available at https://fip-wizard.readthedocs.io/en/latest/ [Last accessed 9 December 2022].

Gundersen, OE and **Kjensmo, S.** 2018. State of the Art: Reproducibility in Artificial Intelligence. *Thirty-Second AAAI Conference on Artificial Intelligence*, 32(1). DOI: https://doi.org/10.1609/aaai.v32i1.11503

Gundersen, OE, Coakley, K and **Kirkpatrick, C.** 2022. Sources of Irreproducibility in Machine Learning: A Review. *arXiv preprint arXiv:2204.07610*.

Jablonka, KM, Patiny, L and **Smit, B.** 2022. Making the collective knowledge of chemistry open and machine actionable. *Nature Chemistry*, 14(4): 365–376. DOI: https://doi.org/10.1038/s41557-022-00910-7

Kirkpatrick, CR, Coakley, K, Schreiber, L, Christopher, J, Katz, D, Stocks, K and **Rao, YD.** 2022. FARR: FAIR in ML, AI Readiness, & Reproducibility Network Postcard. Available in San Diego Supercomputer Center (SDSC) Research Data Services Materials Collection. UC San Diego Library Digital Collections. DOI: https://doi.org/10.6075/J0X92BG3

Law, A. 2022. Fair Implementation Profiles (FIPs) In WorldFAIR: What Have We Learnt? Available at https://worldfair-project.eu/2022/09/30/fips-in-worldfair-what-have-we-learnt-public-workshop-tue-25-october/ [Last accessed 9 December 2022].

LibGuides: Research Data Management: FAIR data. Available at https://ufs.libguides.com/rdm/fair [Last accessed 9 December 2022].

Magnetics Information Consortium (MagIC). Available at https://www2.earthref.org/MagIC/about [Last accessed 9 December 2022].

McCormick, M. 2012. Waterfall vs. Agile methodology. *MPCS, N/A, 3*.

Miksa, T, Walk, P and **Neish, P.** 2020. RDA DMP Common Standard for Machine-actionable Data Management Plans (Version 1.1) [Computer software]. DOI: https://doi.org/10.15497/rda00039

National Institutes of Health. 2020. Final NIH Policy for Data Management and Sharing, 29 October 2020. Available at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html [Last accessed 9 December 2022].

Nicholson, C, Kansa, S, Gupta, N and **Fernandez, R.** 2023. Will It Ever Be FAIR?: Making Archaeological Data Findable, Accessible, Interoperable, and Reusable. *Advances in Archaeological Practice*, 11(1): 63–75. DOI: https://doi.org/10.1017/aap.2022.40

NSF HSI National STEM Resource Hub: Grantsmanship – Training and Resources. Available at https://hsistemhub.org/grantsmanship/ [Last accessed 9 December 2022].

Praetzellis, M and **Riley, B.** 2019. California Digital Library: DMPTool, 8 November 2019. https://cdlib.org/services/uc3/dmptool/ [Last accessed 9 December 2022].

Qualtrics XM – Experience Management Software. Available at https://www.qualtrics.com/ [Last accessed 9 December 2022].

Schultes, E and **Wittenburg, P.** 2019. FAIR Principles and Digital Objects: Accelerating convergence on a data infrastructure. *Data Analytics and Management in Data Intensive Domains: 20th International Conference, DAMDID/RCDL 2018*, Moscow, Russia, October 9–12, 2018, Revised Selected Papers 20 (pp. 3–16). DOI: https://doi.org/10.1007/978-3-030-23584-0_1

Sureshchandra, K and **Shrinivasavadhani, J.** 2008 Moving from waterfall to agile. *Agile 2008 conference*: 97–101. *IEEE*. DOI: https://doi.org/10.1109/Agile.2008.49

Vita, R, Overton, JA, Mungall, CJ, Sette, A and **Peters, B.** 2018. FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability. *Database*, 2018. DOI: https://doi.org/10.1093/database/bax105

Wallis, JC and **Borgman, CL.** 2011. Who is responsible for data? An exploratory study of data authorship, ownership, and responsibility. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1–10. DOI: https://doi.org/10.1002/meet.2011.14504801188

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 160018. DOI: https://doi.org/10.1038/sdata.2016.18

WorldFAIR. Available at https://worldfair-project.eu/ [Last accessed 9 December 2022].

TO CITE THIS ARTICLE:
Kirkpatrick, CR, Coakley, K, Christopher, J and Dutra, I. 2023. Engaging with Researchers and Raising Awareness of FAIR and Open Science through the FAIR+ Implementation Survey Tool (FAIRIST). *Data Science Journal*, 22: 32, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2023-032

Submitted: 16 December 2022
Accepted: 25 May 2023
Published: 06 September 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.