Being FAIR; Having Trust: How clear uncertainty information can increase the accurate reuse of our data.

Shelley Stall[1] and Robert Downs[2]

[1]American Geophysical Union
[2]Columbia University

November 24, 2022

Abstract

Data that are FAIR demonstrate specific characteristics including: ease of discovery, ability to access, community acceptable formats allowing interoperability, and information that supports the decision for reuse. The process used to determine data reuse is commonly called "fit for purpose" or "fit for use". These criteria are defined using relevant factors established by the community for which the data was originally created, and also a "best effort" for criteria needed by other research communities. The FAIR Data Principles support robust documentation of datasets to include the necessary information for reuse. An important part of that documentation, or metadata, is clear documentation of the quality and uncertainty related to the data being considered. When this information is not complete, data has a higher tendency of being used incorrectly leading to inaccurate research, rejected papers, or even retracted papers. The importance of data creators to make their data FAIR – including uncertainty information – directly improves the transparency and integrity of our science today and into the future.

Being FAIR; Having Trust: How Clear Uncertainty Information Can Increase the Accurate Reuse of Our Data

Shelley Stall, American Geophysical Union
Robert R Downs, CIESIN, Columbia University

Data that are FAIR demonstrate specific characteristics including: ease of discovery, ability to access, community acceptable formats allowing interoperability, and information that supports the decision for reuse. The process used to determine data reuse is commonly called "fit for purpose" or "fit for use". These criteria are defined using relevant factors established by the community for which the data was originally created, and also a "best effort" for criteria needed by other research communities. The FAIR Data Principles[1] support robust documentation of datasets to include the necessary information for reuse. An important part of that documentation, or metadata, is clear documentation of the quality and uncertainty related to the data being considered. When this information is not complete, data has a higher tendency of being used incorrectly leading to inaccurate research, rejected papers, or even retracted papers. The importance of data creators to make their data FAIR – including uncertainty information – directly improves the transparency and integrity of our science today and into the future.

FAIR Data are

Findable

Assign persistent IDs (PIDs), provide rich metadata, register in a searchable resource, …

Accessible

Retrievable by their ID using a standard protocol, metadata remain accessible even when data are no longer available…

Interoperable

Use formal, broadly applicable languages, use standard vocabularies, qualified references…

Reusable

Rich, accurate metadata, clear licenses, provenance, use of community standards…

Researchers Can Benefit from FAIR Data:

Findable

Publications must include the data citation that identifies the repository where the data can be accessed. Each data citation must include a persistent identifier that references the landing page where the data can be accessed. Rich data documentation (metadata) enables the use of the data and improves capabilities of potential users to determine whether the data can be useful for meeting their needs. Providing a recommended citation for the data and encouraging users to properly cite the data also can enable those who read about the use of the data to find the data. The recommended citation should include all of the elements of data citation that are described in the Data Citation Guidelines for Earth Science.[2] In particular, the title that is assigned to the data should unambiguously describe the data so that potential users can determine immediately whether the data may be a candidate for their use. Similarly, software that are relevant to the data also must be findable and include these affordances. Metadata about the data and any software should be accessible through relevant catalogs and search engines.

Accessible

The data and data documentation (metadata) are accessible. Online capabilities for using the data should be consistent with the standards and practices of the user community. Such capabilities also should facilitate access to data descriptions, including the metadata (and related documentation and software, if applicable), and offer easy-to-use capabilities for downloading or analyzing the data.

Interoperable

The data to be in a format appropriate for the data domain. Data formats should be considered for dissemination if they are supported by the tools employed by the user community for analyzing data. If multiple formats are used, the repository should have the ability to provide the data in formats that are commonly used by the user community or offer capabilities to obtain the data in such formats. The metadata should employ terminology that facilitates retrieval for relevance and should be encoded in accordance with standards or schemas that are currently utilized by the user community.

Reusable

The data licensing is clear and accessible. The intellectual property rights, including any conditions for using the data, should be described in simple language within the metadata and the data documentation, which also could contain a link to the appropriate licence or legal terminology. Ideally, the data should be as open as possible and free of restrictions.

Key points

As domain communities, we need to encourage our researchers to support FAIR data and include descriptions of quality and uncertainty as part of metadata that is provided with the data. This is critical information for other researchers to assess if a particular data products can be reused for their intended purposes and provide them with information about any conditions for use, and whether such conditions are appropriate for their use.

Well documented, understood data reduces the probability of misuse, and also reduces the likelihood that the data would be used incorrectly. Complete and rich documentation contributes to data transparency and the integrity of the scientific process.

Describing the rights and any conditions for using the data helps to inform potential users whether they may use the data for their intended purposes. Rights declarations should be written in simple, easy-to-understand language, so that potential users can easily decide whether they can use the data. References to particular licenses or usage constraints should be included as a URL to facilitate access by those who need to read such details.

Including a recommended data citation on the data landing page helps to ensure that the data will be cited by those who publish reports on their use of the data. The recommended data citation should contain all of the elements of data citation, including the persistent identifier that references the location of the data landing page.

The metadata that describes the data should be included in data catalogs that are relevant to the community of potential users. Enabling the metadata to be harvested, routinely, by such catalogs will ensure that the metadata are current and distributed to potential user communities.

Fit For Use

An important element to having FAIR data is providing capabilities for determining "Fit for Use" (also known as "Fit for Purpose"). Data products need to include the quality and uncertainty information necessary to determine if data that have been created by others could be valuable for the intended research. Potential users of the data should be able to easily decide whether a particular data product or services is applicable to meet their particular data needs. Such decisions can be supported by describing the data and its use. Furthermore, providing references to studies that have used and cited the data also can enable potential users to determine whether the data are applicable for their intended uses.

The World Data System / Research Data Alliance Assessment of Data Fitness for Use Working Group has recommended five primary categories of dataset fitness for use criteria[3] mapped to the FAIR principles as displayed in parentheses:

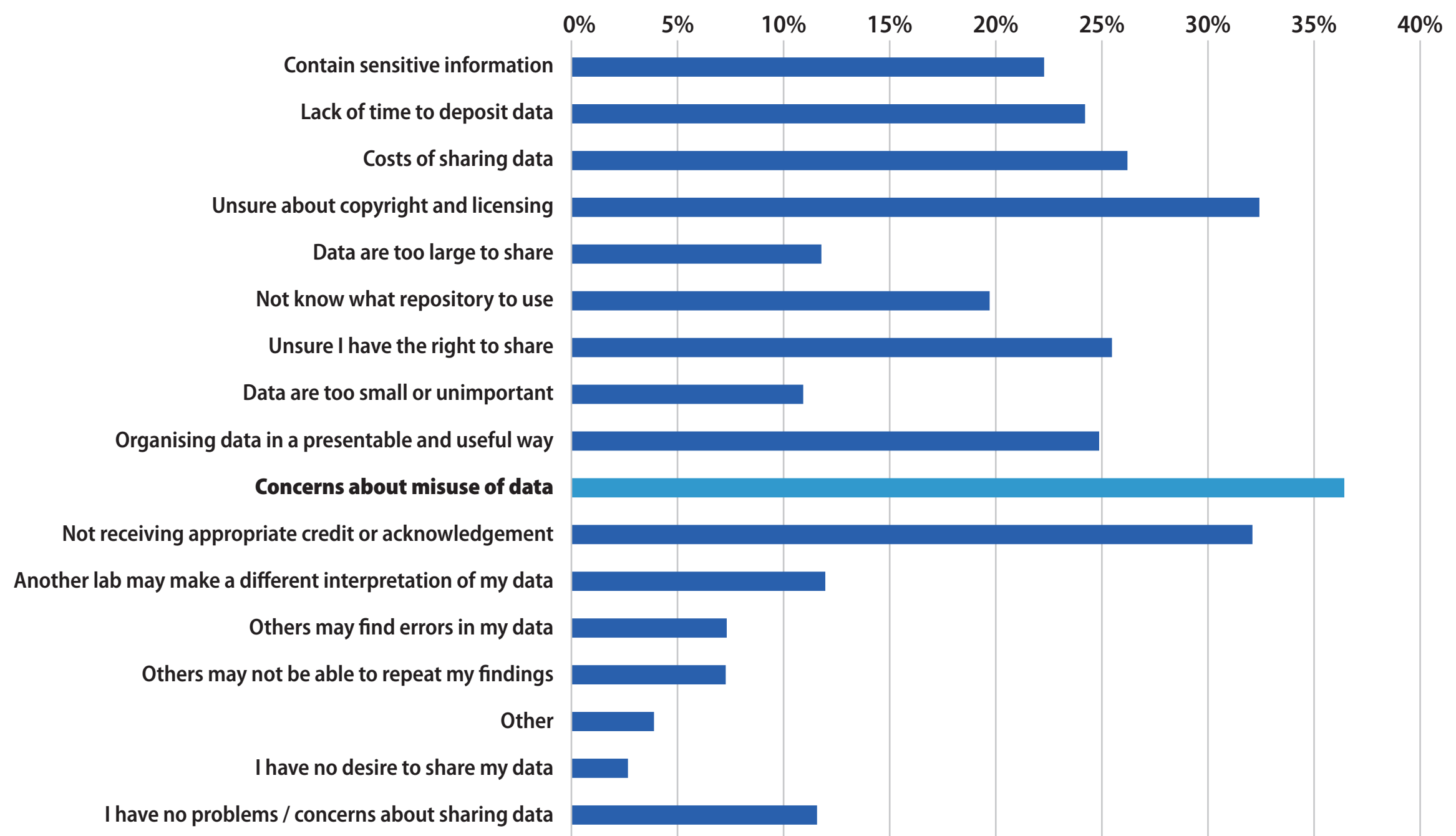
- Metadata completeness (R)
- Accessibility (A)
- Data completeness and correctness (R)
- Findability & interoperability (F, I)
- Curation (leading to overall FAIRness)

There is not a mention of quality and uncertainty in the *Checklist for Evaluation of Dataset Fitness for Use*[4] that describes the criteria to support reuse of the data. It is possible that this information would be defined as part of domain-specific metadata.

Data Misuse

The recent Digital Science Report, The State of Open Data 2019[5] clearly identifies "Concerns about misuse of data" as the primary concern of researchers. Such concerns can be barriers to sharing data for those who might otherwise consider sharing their data with others.

Problems / concerns respondents have with sharing datasets



In Iain Hrynaszkiewicz's commentary "Building Trust to Break Down Barriers within the report, he states:

> **"The biggest barrier to research data sharing and reuse seems to be a matter of trust, and in particular trust in what others may do with researchers' data if it is made openly available."**

Iain Hrynaszkiewicz continues:

> Concerns about data misuse represent a multitude of issues; fears that errors could be found in their work, or that the data could be misinterpreted or research participant privacy be compromised. Researchers might also be concerned that their data will be reused for purposes they did not intend, such as commercial exploitation, or for misleading or inappropriate secondary analyses.[6]
>
> However, trust is more a matter of culture than technology. With repositories being used by around a quarter of researchers[7,8] investing in people rather than infrastructure may be a more pressing issue to change research culture, as Dr Marta Teperek, who coordinates one of the largest institutional data stewardship programmes at TU Delft in the Netherlands, has concluded.[9]

[1] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

[2] ESIP Data Preservation and Stewardship Committee (2019): Data Citation Guidelines for Earth Science Data, Version 2. ESIP. Online resource. https://doi.org/10.6084/m9.figshare.8441816.v1

[3] https://www.rd-alliance.org/group/wdsrda-assessment-data-fitness-use-wg/outcomes/wdsrda-assessment-data-fitness-use-wg-outputs

[4] https://www.rd-alliance.org/system/files/DataFitnessForUse_ChecklistForm_v2_20181218_RDADistribution.pdf

[5] Science, Digital; Fane, Briony; Ayris, Paul; Hahnel, Mark; Hrynaszkiewicz, Iain; Baynes, Grace; et al. (2019): The State of Open Data Report 2019. figshare. Report. https://doi.org/10.6084/m9.figshare.9980783.v2

[6] Wiley Open Science Researcher Survey 2016 [Internet]. [cited 15 Nov 2018]. Available: https:/figshare.com/articles/ Wiley_Open_Science_Researcher_ Survey_2016/4748332/2

[7] Science D, Hahnel M, Treadway J, Fane B, Kiley R, Peters D, et al. The State of Open Data Report 2017. 2017;

[8] Open Data: the researcher perspective - survey and case studies [Internet]. 4 Apr 2017 [cited 15 Nov 2018]. Available: https:/data.mendeley.com/ datasets/bwrnfb4bvh/1

[9] The main obstacles to better research data management and sharing are cultural. But change is in our hands | Impact of Social Sciences [Internet]. [cited 24 Sep 2019]. Available: https:/blogs.[1] lse.ac.uk/ impactofsocialsciences/2018/11/14/ the-main-obstacles-to-better-research- data-management-and-sharing-are- cultural-but-change-is-in-our-hands/