The fair principles: Trusting in fair data repositories

openaccessgovernment.org/article/the-fair-principles-trusting-in-fair-data-repositories/162752

Andy Götz, ESRF data manager and PaNOSC coordinator, discusses the impact of applying the FAIR principles to research data

In the previous article in this series on FAIR data, we explained how the scientific world is undergoing a major change with the widespread adoption of the so-called FAIR principles for research data. FAIR stands for Findable, Accessible, Interoperable and Reusable and was first published in a paper in Nature in 2016.[1]

The FAIR principles were proposed to ensure research data are made available to the scientific community so that they can be found, downloaded, understood, and reused. The goal is to make data used in scientific publications available to the community so they can verify the results, reproduce them, and eventually derive new results from them. Applying the FAIR principles systematically for research data will address the reproducibility – also known as the replicability – crisis [2] in science and make scientific data available for verifying results and used beyond their original purpose.

Making FAIR data the norm

The major European, international, and national scientific organisations, funders and governments have recognised the necessity of making FAIR data the new norm in science and have adopted them. The European Commission is leading projects like the European Open Science Cloud (EOSC).

Scientists are uploading data to well-known data repositories like Zenodo, Figshare, and Dryad to make data available for publications and FAIR. In addition, these data repositories, which serve all communities, research facilities, universities, and scientific communities, play an important role in storing data in self-hosted and managed data repositories. Scientific data repositories almost all aim to be FAIR nowadays. This raises the question of how to know when a data repository is trustworthy and that the data stored there can be trusted to represent what they claim to.
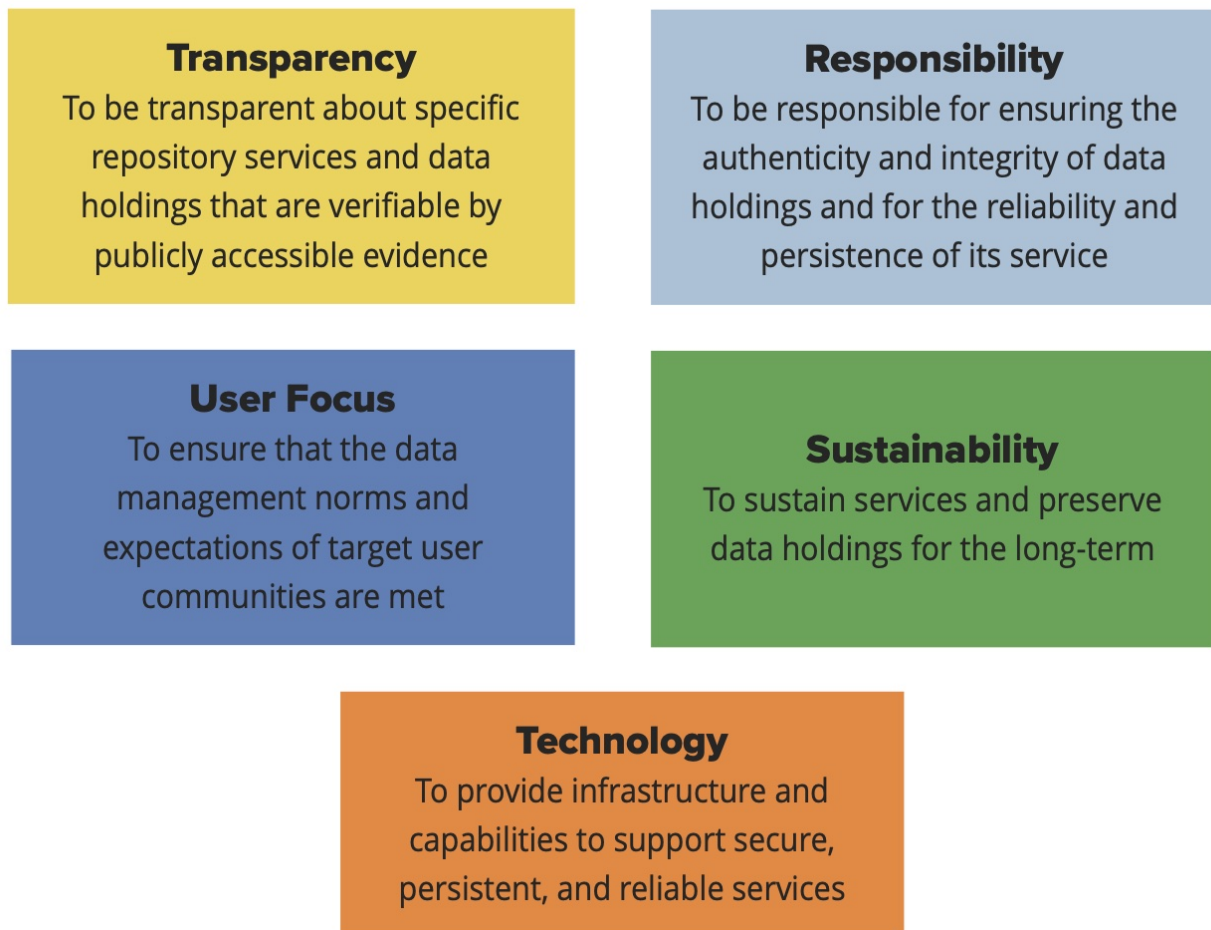
Transparency
To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence

Responsibility
To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service

User Focus
To ensure that the data management norms and expectations of target user communities are met

Sustainability
To sustain services and preserve data holdings for the long-term

Technology
To provide infrastructure and capabilities to support secure, persistent, and reliable services

Figure 1

The TRUST principles

Data are the raw material of scientific evidence. Trusting data downloaded from a data repository depends on the data being FAIR and the repository trustworthy. The Research Data Alliance and a group of data stewards have proposed the TRUST principles for digital repositories [3] (see Figure 1).

The TRUST principles are similar to the FAIR principles in that they are guiding principles and not a set of metrics. The principles address the services which data repositories provide, their transparency and verifiability by the public, the responsibility of data stewards for the data in the repository, the focus on user communities and their metadata standards, the sustainability in the long term and capacity of technical infrastructure to be secure and persistent. The principles help data repository stewards and administrators focus on the quality of their repositories.

A growing number of repositories are going further by certifying their repository to be trustworthy. Three levels of certification can be identified: self-certified with a reviewer approving the self-certification, e.g., CoreTrustSeal; certification by an external body, e.g. Nestor; and ISO certification by an external auditor according to the ISO16363 standard.

The most popular self-certification for scientific data repositories is CoreTrustSeal (CTS). [4] Over one hundred and sixty repositories have been certified to date. This is still a modest number compared to the more than a thousand repositories listed on

Fairsharing.org.

Reasons for this are the extra work involved in certification, the fact that it is not a hard requirement for data repositories (yet) and finally because the added value is not always clear to the repository managers. Nonetheless, going through the certification process is a strong indication of the commitment of the repository managers to provide high-quality data and services for the data they store.

The EOSC offers an opportunity for more repositories to be certified to improve the trust in data repositories. An example of such an approach is the CLARIN ERIC for language resources. All CLARIN data repositories must have CTS certification or have initiated the process to be integrated into the CLARIN.

Another example in the Photon and Neutron community is the ESRF data repository [5], which recently received CTS certification. The ESRF is the first PaNOSC [6] repository to be certified and will hopefully inspire other PaNOSC sites to follow suit. In France, the national plan for Open Science, Ouvrir les Données, [7] proposes support for data repositories to become CTS certified through training, tools, and financing.

The EOSC and the RDA are excellent opportunities to promote trust in data repositories through certifications like CTS and others. We see a grassroots movement towards making data FAIR which could be boosted by building confidence in data repositories through certification. An example of such a grassroots community which serves scientists is the Science Clusters projects which received EOSC funding from the EC and have continued to work together to create a science-driven community. [8]

Can FAIR data increase trust in science?

The verdict is still out on if making data FAIR is enough to restore and increase trust in science. We can say that for data to be accessible over a long period, scientists will require trustworthy data repositories. The increased adoption of certification and the TRUST principles for data repositories are stepping stones towards this goal. We expect that implementing the FAIR principles for data will become a requirement for certified data repositories.

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
2. https://en.wikipedia.org/wiki/Replication_crisis
3. Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7
4. https://www.coretrustseal.org/
5. https://data.esrf.fr
6. https://panosc.eu
7. https://www.ouvrirlascience.fr/home/

8. https://science-clusters.eu

Please Note: This is a Commercial Profile