FAIR data and Net Zero: exploring the interactions

Gabin Kayumbi*, Simon Lambert, Brian Matthews
Scientific Computing Department, STFC Rutherford Appleton Laboratory
Harwell Campus, Didcot OX11 0QX, UK
* Now at The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK
Email gkayumbi@turing.ac.uk, simon.lambert@stfc.ac.uk, brian.matthews@stfc.ac.uk

Abstract

The implications of the push towards FAIR data for progress towards Net Zero carbon emissions in Digital Research Infrastructures (DRIs) have not been much considered. Given the global importance of both these trends, it is certainly worth asking the question "What could be the effect of widespread uptake of FAIR data on the carbon footprint of DRIs?" There might be effects in both directions: easier reuse of data and avoidance of duplication of effort might be countered by the greater power needs of the FAIR data system of resources. The question is complex, but in this paper we present an approach to conceptualising and modelling the interplay between FAIR and Net Zero using Causal Loop Diagrams, and highlight some initial conclusions and open issues.

1. Introduction

The concept of FAIR data (Findable–Accessible–Interoperable–Reusable) sits as a core idea that sharing research data generates opportunities to increase the pace of knowledge discovery and advance scientific progress. It was the topic of a publication in 2016 (Wilkinson et al., 2016), as a result of the workshop "Jointly designing a data FAIRPORT", in 2014, at the Lorentz centre (Leiden, Netherlands). Since their formulation, the principles of FAIR data have gained a great deal of traction and their widespread acceptance will continue to be an important factor in the digital research landscape—for example, the European Open Science Cloud aims to build a "Web of FAIR Data and services for science in Europe"[1], and many national funding bodies have made a commitment to FAIR (Bloemers et al., 2020).

The additional noteworthy development gathering momentum of international significance, is the recognition of the need to reduce carbon emissions to Net Zero[2], and planning for the achievement of that goal. Information technology makes a significant contribution to global carbon emissions through power consumption, and the manufacture and disposal of equipment. In the UK, a wide-ranging study is under way to produce a roadmap for how UK Research and Innovation (UKRI) (the major public research funding body in the UK) can take action to reduce the carbon emissions from data generation, analysis, storage and dissemination from its investments in Digital Research Infrastructure (DRI) towards Net Zero[3].

[1] https://eosc.eu/eosc-about
[2] https://www.un.org/en/climatechange/net-zero-coalition
[3] https://net-zero-dri.ceda.ac.uk/

There is obviously an intersection between these two trends. It is certainly worth asking the question *"What could be the effect of widespread uptake of FAIR data on the carbon footprint of DRIs?"*. Easier findability, lower barriers in accessing repositories, adoption of interoperable formats and greater potential for reuse of research data have the potential to drive data sharing, bring new perspectives, and avoid unnecessary duplication of efforts and resources. However, the process of "FAIRification" of data (that is adding additional information and publishing the data to make it FAIR) requires a complex system of supporting tools and services that in turn require DRIs' resources, such as long-term data preservation, machine processing of data, and assignment and maintenance of persistent identifiers. Whilst embedding FAIR data practices within research is the subject of numerous studies, the balance between costs and benefits of FAIR are yet to be fully understood—and especially in environmental terms.

Within the UKRI Net Zero project already mentioned, and specifically within the sub-project ARINZRIT (Bird, C. *et al., 2023*) a preliminary investigation has been undertaken to explore the relationship between FAIR and Net Zero in the context of DRIs. This paper reports on the progress towards a framework for modelling and understanding the relationships between FAIR and Net Zero in the context of DRIs.

2. Scope and aims

Whilst a considerable amount of work has investigated the environmental impact of ICT (Information and Communications Technology) resources and activities in general (for example Monnin, A., 2019; Royal Society, 2020), discussion of the benefits of making data FAIR highlights work on the opportunities and costs of having or not supporting FAIR data in terms of finances or in terms of scientific innovation (for example Alharbi, E. et al., 2021). There is little consideration of the environmental impact of processes and services involved in making data FAIR. In fact there is a gap in how the specific impact of FAIR in environmental terms is discussed. There is no shortage of evidence for the benefits to the quality and innovation of research itself from data reuse (Pasquetto *et al.*, 2017, CODATA, 2020), but here we are interested in costs and savings in the process rather than the results. There have been studies on the return on investment in financial terms of repositories due to reuse (Beagrie, 2014), and on calculating the break-even point for time spent sharing in a scientific community against time gained by reuse (Pronk, 2018). However these are not related directly to energy use or carbon emissions.

Part of the reason for this gap is surely that analysing the interaction between FAIR and Net Zero is very difficult. It is an interaction between two complex systems, each of which poses its own problems and uncertainties. Expressing the question in the simplest way possible exposes some of the difficulties: "How many megawatt-hours of electricity consumption (or tonnes of CO2) are saved by having this FAIR data repository?" The relevant processes are impossible to model precisely; the scope is unclear (single repository or whole FAIR ecosystem?); much of the interplay is speculative and may be counterfactual (what would have happened if FAIR data had not been available?), and many factors are impossible to quantify.

There are however two ideas that can help at least to frame the question in a more tractable way. One is the idea of *energy proportionality*, defined as "[the objective] to demonstrate that the research design seeks to ensure that the resources used (e.g. hardware purchases, compute time, data storage) will be proportional to the results produced (e.g. outputs, anticipated findings, impacts)" (DHCC Information, Measurement and Practice Action Group, 2022). The concept is also referenced in (Royal Society, 2020) where it is defined as "… whether specific data and computing applications bring environmental or societal benefits that outweigh their own emissions". As intended in this usage, the benefits may be of any kind, without an expectation that these benefits would in any way counteract the emissions. However, it is helpful to adopt a more restricted interpretation, closer to mathematical proportionality of two variables: whether savings in carbon emissions from FAIR data are proportionate to the additional carbon emissions caused by FAIRification. Computation and storage resources will be expended or saved due to the presence and use of FAIR data, and proportionality would mean that the order of saving is the same as or more than the order of growth from resource usage.

The second idea is the avoidance of *resource proliferation* (Monnin, 2019; Royal Society, 2020), through accounting for the environmental impact of devices, justifying the environmental costs of new device purchases, including demonstration of alignment with institutional policies on device recycling.

With this in mind, the aim therefore is to try to understand which contributing factors grow fastest, and which grow more slowly. This is still a highly complex interaction, but even if it is not possible to reach definite conclusions it may be that some pointers can be found to the dominant factors, and that a framework will emerge for examining the relationships and their implications as a basis for further discussion and analysis.

3. Methodology

Expanding on the aforementioned challenges in exploring the environmental impact of FAIR data, we propose that by conceptualising the processes underpinning FAIR in relation to the digital activities undertaken, we can provide a framework for analysing the problem by integrating the concepts of energy proportionality and resource proliferation. In particular, we consider among the processes that support FAIR those linked to energy consumption in ICT resources. Those processes are related to FAIR data storage and data management: long-term preservation, FAIR-oriented curation, trusted repositories, metadata creation, persistent identifiers, data reuse, sharing, quality assurance, data integration and transfer. We examine the different factors at play by looking at those processes as part of a dynamic system in which elements positively or negatively influence each other. In modelling this system, we set the generation of research data as the initiator and carbon emissions as the sink that represents the resulting environmental impact.

We do not however intend this dynamic system to be a comprehensive description that models the cost and benefits of FAIR data in terms of environmental impact. Nor do we present an exhaustive list of factors at play and the magnitude of respective influences. Instead, we set a framework to discuss the problem by revealing key elements and formulating the questions to consider in order to evaluate the environmental impact of FAIR.

To describe the dynamic system of FAIR processes and how its different components influence each other, we build a Causal Loop Diagram (CLD). CLDs are a tool in Systems Thinking, used to explore relationships between different components of a dynamic system (Meadows, 2015;, Haraldsson (2004). This method sits halfway between the qualitative and quantitative spectrum of systems mapping methods. It is particularly useful in capturing and modelling relationships between entities of a dynamic system in a qualitative way without quantifying the influences. In our approach, CLDs are built by representing nodes as FAIR processes and data activities identified in previous works

(Stephens, A. *et al.*, ,2023; Bird, C. *et al., 2023)*]. We used the free software *Loopy*[4] to build our CLD. Whilst CLDs are often used to display influences between elements of a compound system, they are particularly useful in revealing loops—either reinforcing or balancing loop— which are areas requiring major focus. We will extract some of those loops in our discussions.

4. Building a framework using CLDs

In this section, we construct a CLD by identifying the most relevant processes that underpin making data FAIR. We aim at highlighting influences, in alignment or tension (positives and negatives) between underlying FAIR-related processes and activities that ultimately have an impact on carbon emissions, in particular when examined under the light of energy proportionality and resource proliferation. In these regards, reinforcing and balancing loops revealed by the CLD are of particular importance. They can potentially lead to intervention points to consider when seeking to mitigate the environmental impact of FAIR.

Based on the knowledge built in the course of the FAIRsFAIR project[5] and our recent contribution to the Net Zero and ARINZRIT projects, we select and colour code the following relevant processes as nodes of a FAIR-Net Zero CLD:

- Initialiser (red node)
  - DATA GENERATION
- FAIR oriented activities (blue nodes) - use of:
  - FAIR-ORIENTED CURATION; METADATA CREATION; PERSISTENT IDENTIFIERS; LONG-TERM PRESERVATION; DATA REUSE
- Storage activities (yellow nodes) - use of:
  - TRUSTED REPOSITORIES; REMOTE SERVERS
- Data processing & transfer activities (brown nodes) - use of:
  - ARTIFICIAL INTELLIGENCE / MACHINE LEARNING (AI/ML); DATA QUALITY; DATA INTEGRATION; DATA SHARING; DATA TRANSFER
- Sink (green):
  - CARBON EMISSIONS

Figure 1 shows a CLD where the nodes represent the processes and activities involved in FAIR as listed above. The (+) arrow indicates a positive reinforcement, that is as the process/activity indicated in the node changes, the one in the next node changes in the same direction. Conversely, (-) points at a change in the opposite direction. The CLD may be explored at http://bit.ly/3FOuuj5 .

[4] https://ncase.me/loopy/

[5] https://www.fairsfair.eu/

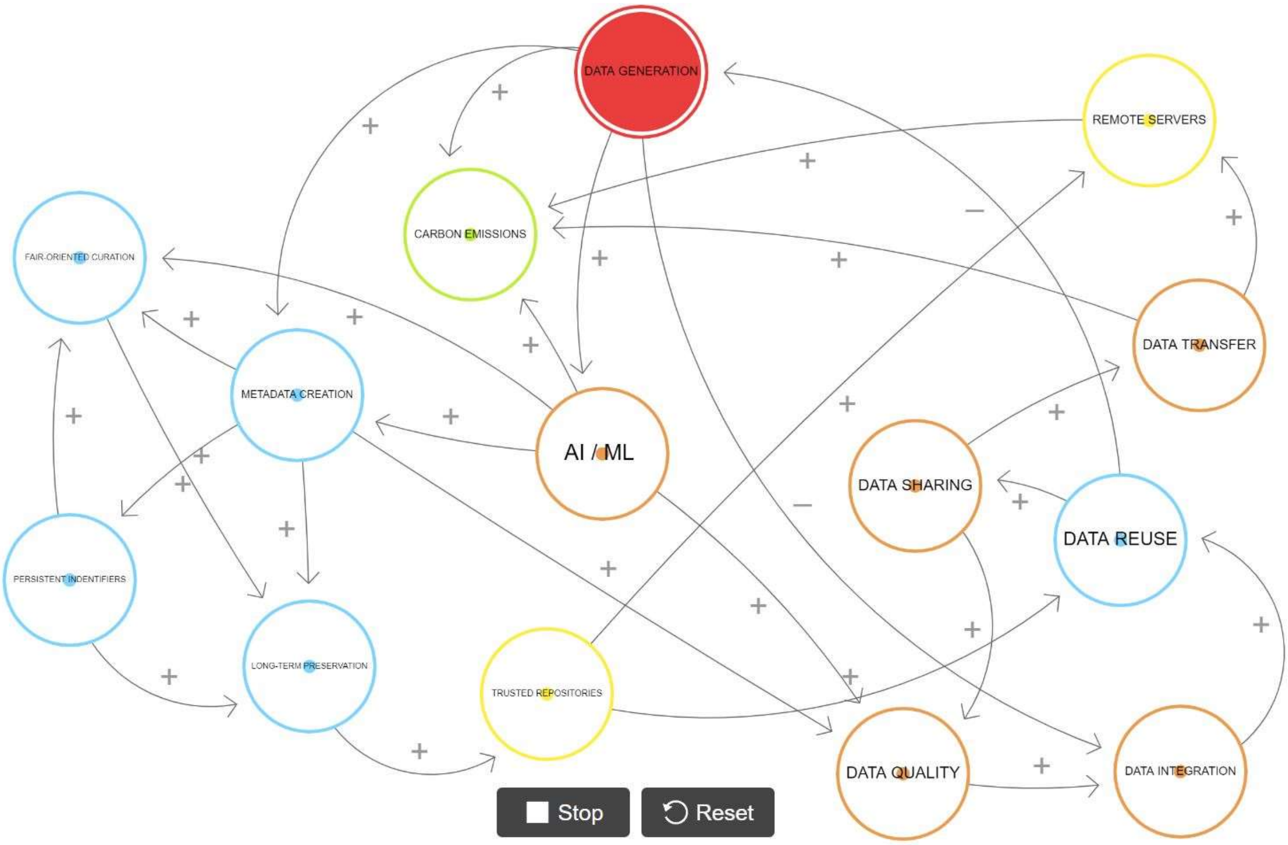

Figure1. Causal Loop Diagram representing the influences (positive or negative) between FAIR processes and data activities and their impact on carbon emissions..

The following key points are derived from the CLD in Fig.1. The potential environmental impact of FAIR data is linked to the carbon footprint of data storage, transfer and communication processes. Sharing data across different locations and networks requires a significant amount of energy, particularly when using high-speed networks and data-intensive applications. This energy consumption can potentially lead to higher carbon emissions. Whilst this is the case for data sharing in general, we want to look at the specifics of FAIR processes. Figure 1 CLD reveals a balanced loop that is extracted in Fig.2 : DATA GENERATION + METADATA CREATION + DATA QUALITY + DATA INTEGRATION + DATA REUSE - DATA GENERATION. The CLD may be explored at http://bit.ly/3JM1c6a. More metadata is created as a consequence of newly generated research data, and that enhances the quality of data which facilitate its integration and ultimately reuse. Furthermore, the reuse of data reduces the need for more research data to be generated, and consequently carbon emissions are reduced too. However, data reuse also leads to more data being shared, then more data being transferred and ultimately more carbon emissions. How can those two loops be interpreted in terms of the carbon emissions? Which one is the dominant loop over time? Those are questions that can be analysed when we integrate the energy proportionality related to the loops and the resource proliferation associated to the ICT involved in the processes. To account for the environmental impact, making data FAIR should consider finding a balance between the resources used in each one of the specific processes and the results expected in output.

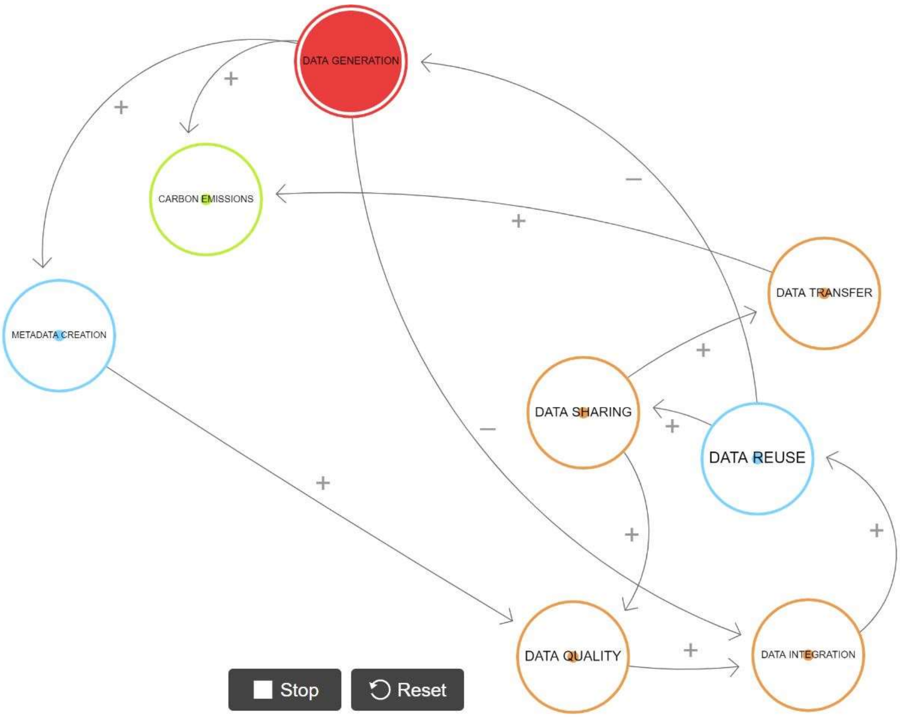

Figure 2. Balancing loop (DATA GENERATION + METADATA CREATION + DATA QUALITY + DATA INTEGRATION + DATA REUSE - DATA GENERATION) extracted from the CLD in Fig.1.

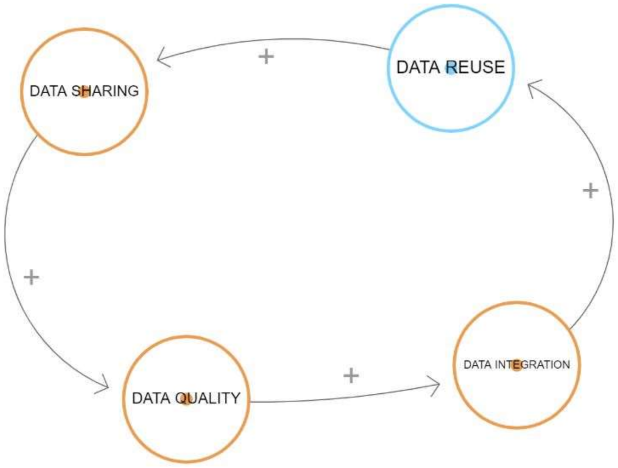

Figure 3. Reinforcing loop (DATA REUSE + DATA SHARING +DATA QUALITY + DATA INTEGRATION) extracted from the CLD in Fig.1.

Figure 3 instead shows a reinforcing loop (DATA REUSE + DATA SHARING + DATA QUALITY + DATA INTEGRATION) in which the more data is reused, the more it is shared which improves its quality and facilitates integration from multiple sources which again lead to more reuse and so forth. This is an example where FAIR processes generate a cycle of influences that go on increasing over time. The CLD may be explored at http://bit.ly/3TJ7Xu2 . *What is the energy associated with the resources involved in this case?* CLD in Fig.1 reveals that loops that include DATA INTEGRATION and DATA REUSE nodes require closer analysis of influences in the processes involved. A legitimate query could therefore be: *"To what extent accounting for energy proportionality and resource proliferation within each node those loops can balance the carbon emissions in this loop?".*

Following the CLDs description of influences and impact of processes at play in FAIR, and the consideration of energy proportionality and resource proliferation in the framework of discussion on the impact of FAIR, we identify some potential intervention points below. They represent actions that can be taken to mitigate the environmental impact of FAIR:

- Accounting for the full environmental impact of ICT resources: FAIR long-term preservation among other processes, should consider the embodied carbon footprint involved. For example, by conducting the associated life-cycle assessment (LCA)[6].
- Long-term preservation: a tiered approach should be considered where possible. One strategy could be to allocate resource-intensive repositories for metadata whilst assigning nearline or offline repositories for data. Data could then be made available upon request.
- Trusted repositories: as a corollary of above point, adjust the repository technologies for access by introducing a tiered repository system that supports reducing the environmental impact of digital delivery.
- AI/ML: although their adoption can lead to resource-intensive applications, practitioners can leverage AI/ML techniques to provide powerful tools in achieving efficiency in FAIR data management.
- More generally, FAIR Data stewardship should adopt standards that reflect the focus on striving for optimal management in the context of Net Zero.

5. Conclusions and future work

Whilst it is still a fairly recent development, the amount of FAIR data is significant and growing rapidly, and the research community has an obligation to critically evaluate FAIR practices in relation to its environmental impact. The whole research community, and primarily researchers themselves, are responsible for the environmental impact of FAIR processes regardless of its magnitude and quantifiability. Moving toward environmentally sustainable FAIR processes requires critically examining the underlying assumptions and activities that shape current practice.

Notwithstanding the difficulty in quantifying its related costs and benefits, we have started to describe a framework conducive to understanding the environmental impact of FAIR. However, there is still work to do in defining the full extension of our framework. That would include: extracting more mediating factors between nodes; explicating variations over time of quantities represented in the nodes; understanding of the differences between long and short-term consequences of influences.

Additionally, some questions are left open: with what granularity to account for energy proportionality and resource proliferation in making data FAIR? What metrics to use to express the integration of energy proportionality and resource proliferation in measuring the environmental impact of FAIR?

We believe that discussions and potential answers to those questions will require a breakdown and careful analysis of each step in making data FAIR combined with the research work being conducted in green computing.

Acknowledgements:
This work was supported by the UK Research and Innovation Net Zero Digital Research Infrastructure Scoping (NE/W007134/1); funded by the Natural Environment Research Council (NERC) and managed through the Centre for Environmental Data Analysis (CEDA)..
We are grateful to Caroline Bird, Chris Preist, Carolynne Lord, Adrian Friday, Kelly Widdicks and Adrian Jackson from the ARINZRIT team for their insights.

[6] LCA is a process codified by the International Organization for standardisation as ISO 14040.

6. References

Alharbi, E., Skeva, R., Juty, N., Jay, C., & Goble, C. (2021). Exploring the current practices, costs and benefits of FAIR Implementation in pharmaceutical Research and Development: A Qualitative Interview Study. Data Intelligence, 3(4), 507-527. https://doi.org/10.1162/dint_a_00109.

Beagrie, N. & Houghton, J. 2014. The Value and Impact of Data Sharing and Curation. A synthesis of three recent studies of UK research data centres. Jisc report. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Bird, C., Preist, C., Lord, C., Friday, A., Widdicks, K., Jackson, A., Kayumbi Kabeya, G., & Lambert, S. (2023). Applying Responsible Innovation to the Net Zero Research Infrastructure Transformation (ARINZRIT): Summary Briefing. Zenodo. https://doi.org/10.5281/zenodo.7689198

Bloemers, M. & Montesanti, A. The FAIR Funding Model: Providing a Framework for Research Funders to Drive the Transition toward FAIR Data Management and Stewardship Practices. Data Intelligence 2020; 2 (1-2): 171–180. https://doi.org/10.1162/dint_a_00039

CODATA Coordinated Expert Group (2020). Open Science for a Global Transformation: CODATA coordinated submission to the UNESCO Open Science Consultation. Zenodo. https://doi.org/10.5281/zenodo.3935461

DHCC Information, Measurement and Practice Action Group. (2022). A Researcher Guide to Writing a Climate Justice Oriented Data Management Plan (v0.6). Zenodo. https://doi.org/10.5281/zenodo.6451499

Haraldsson, H. (2004) Introduction to System Thinking and Causal Loop Diagrams. Reports in ecology and environmental engineering, KFS AB, Lund, Sweden

Meadows, D. H. (2015). Thinking in Systems. Chelsea Green Publishing.

Monnin, A. (2019). Lean ICT: Towards digital sobriety. The Shift Project technical report. https://theshiftproject.org/en/lean-ict-2/

Pasquetto, I V, Randles, B M and Borgman, C L 2017 On the Reuse of Scientific Data. Data Science Journal, 16: 8, pp. 1–9, DOI: https://doi.org/10.5334/dsj-2017-008\.On the Reuse of Scientific Data. Available from:https://www.researchgate.net/publication/315547183_On_the_Reuse_of_Scientific_Data.

Pronk, T.E. 2019 The Time Efficiency Gain in Sharing and Reuse of Research Data, http://doi.org/10.5334/dsj-2019-010

Royal Society: Digital technology and the planet: Harnessing computing to achieve net zero (2020) https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/

Stephens, A., Kayumbi, G., Lambert, S. (2023). UKRI Digital Research Infrastructure Mapping Survey Dataset (for Net Zero Scoping Project) (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7784877

Van Vlijmen, H., Mons A., Waalkens, A., Franke, W., Baak, A., Ruiter, G., Neefs, J-M. The need of Industry to go FAIR. Data Intelligence 2(2020), 276–284. https://doi.org/10.1162/dint_a_00050.

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