# Increasing Reproducibility in Science by Interlinking Semantic Artifact Descriptions in a Knowledge Graph

Hassan Hussein[1][0000−0003−3975−5374], Kheir Eddine Farfar[1][0000−0002−0366−4596], Allard Oelen[1][0000−0001−9924−9153], Oliver Karras[1][0000−0001−5336−6899], and Sören Auer[1,2][0000−0002−0698−2864]

[1] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{hassan.hussein,kheir.farfar,allard.oelen,oliver.karras,soeren.auer}@tib.eu
[2] L3S Research Center, Leibniz University of Hannover, Hannover, Germany

**Abstract.** One of the pillars of the scientific method is reproducibility – the ability to replicate the results of a prior study if the same procedures are followed. A lack of reproducibility can lead to wasted resources, false conclusions, and a loss of public trust in science. Ensuring reproducibility is challenging due to the heterogeneity of the methods used in different fields of science. In this article, we present an approach for increasing the reproducibility of research results, by semantically describing and interlinking relevant artifacts such as data, software scripts or simulations in a knowledge graph. In order to ensure the flexibility to adapt the approach to different fields of science, we devise a template model, which allows defining typical descriptions required to increase reproducibility of a certain type of study. We provide a scoring model for gradually assessing the reproducibility of a certain study based on the templates and provide a knowledge graph infrastructure for curating reproducibility descriptions along with semantic research contribution descriptions. We demonstrate the feasibility of our approach with an example in data science.

**Keywords:** Reproducibility Assessment, Scholarly Knowledge Graph, FAIR Data Principles

## 1 Introduction

One of the guiding principles for scientific work is to guarantee the reproducibility of research findings as long as the researcher employs the same methodology as the original study. Reproducibility is a major concept in which we distinguish and describe access to scientific resources and their completeness to the extent necessary to efficiently and successfully engage with scientific research [8]. Due to the variety of methodologies used across different research fields, ensuring reproducibility is a challenging issue. Employing subjective processes, such as visual interpretation or data analysis, can result in diverse outcomes even when the exact methods are used, thus deepening the problem. Reproducibility issues can result in erroneous results, and a decline in public confidence in science. To assess the validity and reliability of results, it is crucial to be able to reproduce a study's results using the same methodology. So, guaranteeing reproducibility in scientific research can be challenging and complex. The FAIR data principles by Wilkinson et al. [27] are one of the most widely used guidelines for increasing machines' ability to automatically reuse data (i.e., machine actionability). The FAIR principles provided a conceptual model for outlining our novel approach , which increases the reproducibility of research results by semantically denoting and connecting all relevant artifacts, such as data, or software scripts via a knowledge graph. In this work, we address the following research questions: *What are the requirements for creating a general reproducibility assessment for various scientific fields?*, and *How to foster collaboration and knowledge exchange in scientific communities?* In this study, we leverage the the Open Research Knowledge Graph (ORKG[1]) infrastructure to select some use case and to implement our reproducibility score. For answering the research questions, we design a semantic template model for knowledge graphs that enables the construction of standard descriptions needed to increase the reproducibility in different disciplines. Additionally, we develop a scoring model to gradually assess the reproducibility of a study based on templates. Based on our case study, we are confident that semantic templates will help researchers describe research artifacts relevant to improve the reproducibility of their work. The article is structured as follows. In section 2, we discuss related work. In section 3, we present our proposed scoring pillars. In section 4, we explain how we implemented the templates and the scoring models. In section 5, we present a use case based on our approach. Finally, in section 6, we conclude and discuss potential future work.

## 2 Related Work

Recent studies have shown that reproducibility is a significant issue in the scientific community. According to a survey conducted by the journal Nature, more than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than 50% have failed to reproduce their own experiments [1]. This lack of reproducibility can lead to wasted resources, and false conclusions. Reproducibility faces some major issues, including setting up the proper technological infrastructure [6], the need to encourage and motivate researchers to publish and disclose their work publicly [1,7], and promoting the best approaches among researchers [13,2]. Howison et al. [12] emphasized that the challenges of sharing code and gaining academic credit for open-source collaborations have not yet been adequately addressed. There are some convincing proofs that much scientific research has not been able to be replicated [11]. In our literature review, we could find some common problems that hinder the research reproducibility as follow:

[1] `https://orkg.org/`

- `The unavailability of the replication data`: Reproducibility is a tribulation in both the natural and social sciences [10,18,21]. Two-thirds of all political science publications disclosed in the American Political Science Review (APSR) between 2013 and 2014 did not furnish replication materials, according to Key [15]. In genetics [20], medicine [24], economics [3,5,17], and sociology [19] is still the same problem. As per Vines et al. [26], the availability of research data declined dramatically over time following publication, by 17% annually in 516 studies with article ages ranging from 2 to 22 years.
- `The data accessibility`: Data accessibility is a crucial aspect of scientific research that can have a significant impact on the reproducibility of studies [22]. Furthermore, a comparison on ORKG[2] shows absolutely inadequate percentage of data accessibility in various research domains (e.g, agricultural science..,etc) across different geographical regions ranging from 3.26% to 39.26%.
- `The data completeness`: Chen et al. [4] argue that just concentrating on the data is insufficient. They stress that the data must be accompanied by software, workflows, and explanations. These elements must all be documented throughout the typical iterative and completed research lifecycle to be prepared for a timely open release of the results.
- `License`: Feger and Wozniak [8] state that external variables such as licensing limitations and expiration periods may impact sharing.

Another comparison on the ORKG[3] demonstrates various approaches for computational reproducibility. The comparison originated from work [9] that discusses most of the reproducibility approaches, specifically sharing data and code, etc. The work also presented some of the tools used in each approach (e.g., the TIER Protocol, Do-Files, etc.). After reviewing some of these approaches and tools to comprehend the issues underlying each one of them. We can conclude that most of these approaches are time-consuming to implement, have a steep learning curve for new users, have compatibility issues with some software or systems, and user interface problems.

## 3 FAIR-based Reproducibility

Our model will foucs on the four pillars: availability, accessibility, linkability, and license. These pillars directly impact the reproducibility of artifacts in any study. We now discuss each of the pillars and what they mean in detail.

- **Availability:** We define availability as the willingness of researchers to make their artifacts, such as data, resources, and methods, voluntarily available to other scientists.
- **Accessibility:** In our earlier study [14] we set the accessibility measures on a top maturity level when assessing the accessibility of a knowledge graph

[2] `https://orkg.org/comparison/R589387/`
[3] `https://orkg.org/comparison/R589371/`

(KG). In addition, article 06.1 in data access and research transparency (DA-RT)[4] states that "researchers making evidence-based knowledge claims should reference the data they used to make those claims." The guide further declares that "if these are data [the researchers] themselves generated or collected, researchers should provide access to those data or explain why they cannot."
- **Linkability:** We suggest that using ontologies to link scientific data with other sources is crucial for facilitating the reproducibility of scientific findings. The World Wide Web Consortium (W3C)[5] has been working to develop standards for linked data[6] to facilitate the integration of data from different sources.
- **License:** A valid license makes it feasible for researchers to comprehend the responsibilities and constraints associated with using the artifact, ensuring that they can use, and modify the artifact legally as required. Reproducibility was found to be negatively correlated with the lack of a license [23].

## 4 Implementation

The ORKG infrastructure implements best practices, such as the FAIR principles, offers a wide range of services to make it easier to curate, share, and use FAIR scientific information [25]. Because of these features, the ORKG is an ideal infrastructure to implement our approach. In this section, we explain what the reproducibility score is and how it works. In addition, we present the ORKG template engine and how it supports implementing the reproducibility score.

### 4.1 Reproducibility Score

The reproducibility score is a method aiming to evaluate some of the fundamental requirements (e.g., availability) for making the scientific contribution or experiment reproducible. As illustrated in Figure 1, the first stage is allowing the user to add their paper's metadata. Secondly, the user picks the appropriate template for their research contribution or constructs a template if needed. (Figure 2 illustrates a template). We propose to add a checkbox to mark a certain property in a template as reproducibility description property, thus distinguishing between conventional templates and templates that users can employ specifically for reproducibility. Thirdly, the user describes the research contribution guided by the chosen template. Finally, the system automatically computes the reproducibility score for the given contribution.

Given a reproducibility score, the user can review the full report as shown in Figure 3. By availability we mean that the contribution curator utilized the resource by a value. The accessibility attribute represents a resource of type URL(ORKG has different data types including resource, URL...etc) and has an

[4] https://www.dartstatement.org/2012-apsa-ethics-guide-changes

[5] https://www.w3.org/

[6] https://www.w3.org/standards/semanticweb/data

HTTP response code of 200, which denotes that the related artifact is reachable. The system considers this resource unsuitable for this test if it is not of type URL. Linkability indicates that the data curator linked the resources to a reliable ontology. Finally, a URL-based resource that is available under an open license, such as the MIT License, is considered to have a favorable License, allowing for reuse and redistribution. Figure 3.

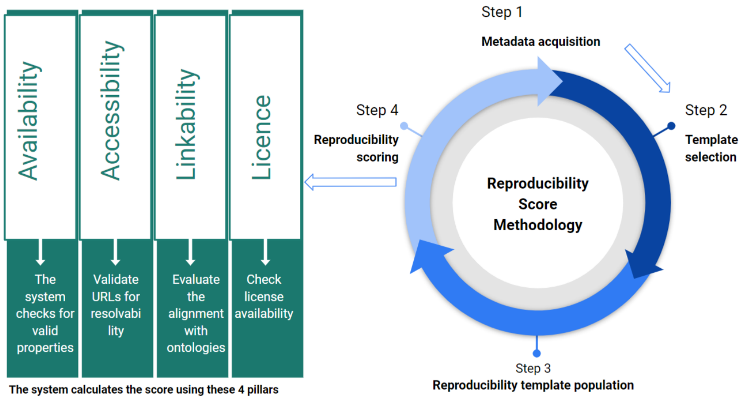

Availability
Accessibility
Linkability
Licence
The system checks for valid properties
Validate URLs for resolvability
Evaluate the alignment with ontologies
Check license availability
The system calculates the score using these 4 pillars
Step 1
Metadata acquisition
Step 2
Template selection
Step 3
Reproducibility template population
Step 4
Reproducibility scoring
Reproducibility Score Methodology

Fig. 1: Workflow illustrating the architecture and components of the reproducibility score, showcasing the systematic process for assessing reproducibility in scientific research.

## 4.2 Score Calculation

The score calculation assesses the four pillars, namely the availability, accessibility, linkability, and licenses of artifacts automatically once the end user adds their contribution. A higher reproducibility score implies a higher degree of allowing other researchers to repeat the given experiment. The artifacts linked in the reproducibility template are categorized into one of three cases:

- **Valid**: the artifact meets the criteria for being complete, accurate, and relevant to the problem at hand.
- **Inapplicable**: an artifact cannot be used in a specific context or for a specific purpose due to its limitations or characteristics. For example, if the data is not of type URL, it cannot be checked for accessibility.
- **Not Valid**: data is incomplete, inaccurate, or does not fulfill the standards for quality and reliability. For example, if data is not linked with a trusted ontology when its type is a resource then it is not valid.

The score for the four pillars is the trimmed mean for set of properties for a given contribution of a matrix with 4 columns(pillars) and n rows(properties values), while excluding the highest and lowest values to avoid outliers is calculated as follows:

$$\overline{X}tr = \frac{1}{n-2}\sum i = 2^{n-1}\operatorname{trim}\left(x_i, 0.5\right) \tag{1}$$

where $\operatorname{trim}\left(x_i, 0.5\right)$ represents the trimming of $x_i$ by 0.5 (i.e., removing the highest and lowest 50% of values) to avoid outliers.

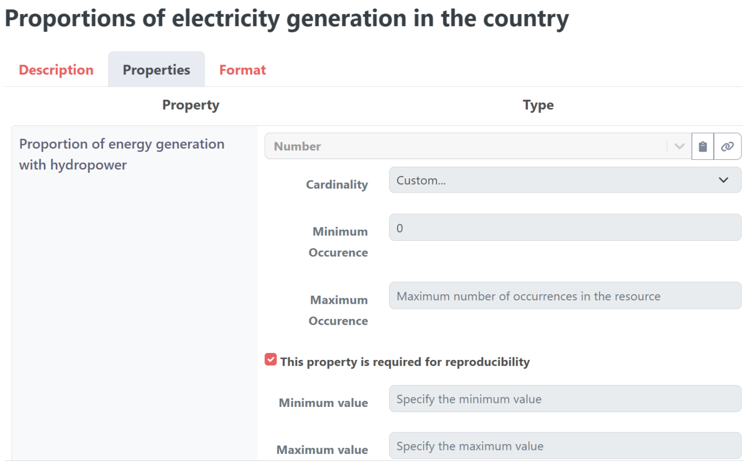
To calculate the trimmed mean of each row, you can apply this equation to each row of the matrix separately. To calculate the trimmed mean of each column, you can transpose the matrix and apply the equation to each row (which corresponds to each column of the original matrix).

$$\overline{X}_{tr,1} = \frac{0+100}{2} = 50 \tag{2}$$

We repeat this process for each of the other rows, and obtain the following trimmed means. The system continuously allows the user to improve their contribution's reproducibility score. As the user can preserve editing their contribution and add the missing artifacts as needed.

### 4.3 ORKG Templates

The ORKG provides a template system that empowers domain experts to define the structure of contributions. Templates automatically generate user-friendly input forms to assist researchers in providing the necessary data. The system



Proportions of electricity generation in the country

| Property | Type |
|---|---|
| Proportion of energy generation with hydropower | Number |

Cardinality: Custom...

Minimum Occurence: 0

Maximum Occurence: Maximum number of occurrences in the resource

☑ This property is required for reproducibility

Minimum value: Specify the minimum value

Maximum value: Specify the maximum value

Fig. 2: A screenshot of the ORKG template system, depicting a property, property type, and cardinality.

enables the definition of constraints on properties, such as data types and cardinality of values, to generate appropriate input forms and perform data validation.

It implements a subset of the Shapes Constraint Language (SHACL) [16] which proportionately enables the interoperability with existing SHACL-based systems and ensures uniformity in the verification and management of data. The template system serves as a controller, supervising what data should be collected based on the research field and problem. Furthermore, it aligns with the FAIR data principles. By utilizing external ontologies and linking data to them, researchers are able to increase the discoverability, accessibility, and reusability of their contributions. By using a template system, researchers can compare their work with others and establish a common language (i.e., standardized vocabulary). In addition, to further promote and incentivize reproducibility within the ORKG platform, we propose the implementation of a new feature: a "required for reproducibility" checkbox. When the template designer selects this checkbox, it implies that the associated property is considered a crucial element in determining the overall reproducibility score of a given artifact.

## 5 Reproducibility Score Use Cases

In this section, we demonstrate a use case for the proposed reproducibility score in data science. As shown in Figure 3, the horizontal axis of the report shows the average reproducibility score for a given property, while the vertical axis represents the unique score for each of the four pillars of reproducibility.

### 5.1 Papers With Code

The Papers With Code (PWC)[7] system aims to promote information sharing about the evolution of machine learning research. A subset of this data is fed to ORKG by modeling the different aspects covered by the data in a structured way. This includes the algorithms used, their evaluation on specified benchmark datasets, and the metrics used to measure performance (e.g., precision, recall, f-measure). An ORKG template named Leaderboard [8] is created to guide users in adding more data to the graph, adhering to the established model. We chose one of the papers[9] imported into ORKG as an example to demonstrate the score concept (Figure 3). We used this paper because it shows some properties that use URLs. Next we explain what this report means:

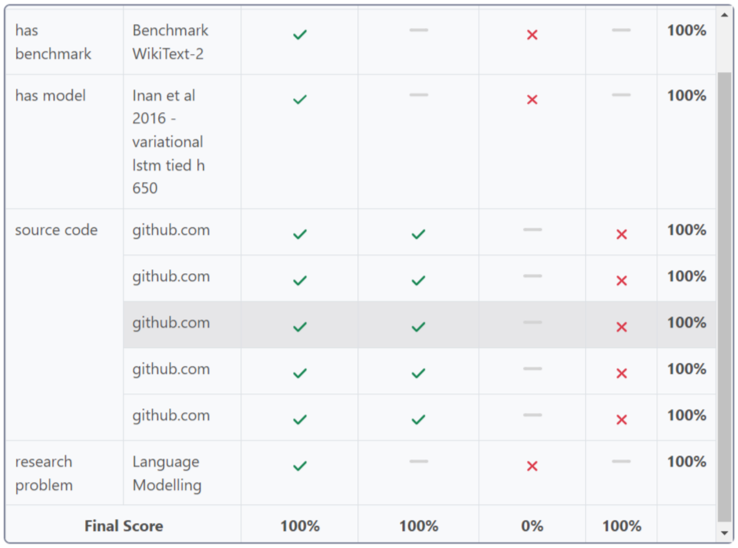
- **Availability**: All data is available so that the vertically-aggregated score is 100%.
- **Accessibility**: The score is 100% for some properties that are of type URL (e.g., source code). The system also checks to see if the specified URL can be reached with an HTTP response code of 200. In addition, all the other resources that are not of type URL (e.g., has model) were identified as inapplicable for this test. As a result, the vertically-aggregated score is 100%.

[7] `https://paperswithcode.com/about`

[8] `https://orkg.org/template/R107801`

[9] `https://orkg.org/paper/R478126`

R



| | | | | | | |
|---|---|---|---|---|---|---|
| has benchmark | Benchmark WikiText-2 | ✓ | — | ✗ | — | 100% |
| has model | Inan et al 2016 - variational lstm tied h 650 | ✓ | — | ✗ | — | 100% |
| source code | github.com | ✓ | ✓ | — | ✗ | 100% |
| | github.com | ✓ | ✓ | — | ✗ | 100% |
| | github.com | ✓ | ✓ | — | ✗ | 100% |
| | github.com | ✓ | ✓ | — | ✗ | 100% |
| | github.com | ✓ | ✓ | — | ✗ | 100% |
| research problem | Language Modelling | ✓ | — | ✗ | — | 100% |
| **Final Score** | | **100%** | **100%** | **0%** | **100%** | |

Fig. 3: A view of how the reproducibility score computed for the PWC use case.

- **Linkability**: The properties “has model” and “research problem” are of the type resource but not associated with ontologies. For this reason, the system scored them 0%.
- **License**: The system assigned a score of 100% to properties “has benchmark” and “has model” because they are not of type URL and is inapplicable for this test. Furthermore, the system allocated a 0% score for the property ”source code” as it is of type URL and the system could not correlate a proper license for the given URLs.

## 6 Conclusion and Future Work

The approach focuses only on measuring the four pillars: availability, accessibility, linkability, and license, to provide a thorough assessment of the dependability and robustness of a study’s findings. Our use of ORKG templates in this score implementation permits an effective and automated calculation of the score. To show the adaptability and effectiveness of this approach in different domains we presented, a use case in data science. In future work, we plan to implement the reproducibility score as an open assessment tool. Publishers (e.g., Springer) can integrate it into their services. We also intend to do a qualitative assessment for the reproducibility score.

*Supplemental Material Statement:* Source code is available on Github[10].

[10] `https://gitlab.com/TIBHannover/orkg/orkg-frontend/-/merge_requests/1015`

## References

1. Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015). https://doi.org/10.1126/science.aac4716, https://www.science.org/doi/abs/10.1126/science.aac4716
2. Buys, C.M., Shaw, P.L.: Data management practices across an institution: Survey and report **3**(2), 1225 (2015). https://doi.org/10.7710/2162-3309.1225
3. Chang, A.C., Li, P.: Is economics research replicable? sixty published papers from thirteen journals say ”usually not” (2015), https://shorturl.at/jlpxQ
4. Chen, X.: Open is not enough **15**, 7 (2019)
5. Dewald, W.G., Thursby, J.G., Anderson, R.G.: Replication in empirical economics: The journal of money, credit and banking project **76**(4), 587–603 (1986), https://www.jstor.org/stable/1806061, publisher: American Economic Association
6. Feger, S.S., Dallmeier-Tiessen, S., Woźniak, P.W., Schmidt, A.: The role of hci in reproducible science: Understanding, supporting and motivating core practices. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–6. CHI EA '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290607.3312905, https://doi.org/10.1145/3290607.3312905
7. Feger, S.S.: Interactive tools for reproducible science - understanding, supporting, and motivating reproducible science practices p. 221 (2020)
8. Feger, S.S., Woźniak, P.W.: Reproducibility: A researcher-centered definition **6**(2), 17 (2022). https://doi.org/10.3390/mti6020017, https://www.mdpi.com/2414-4088/6/2/17
9. Figueiredo Filho, D., Lins, R., Domingos, A., Janz, N., Silva, L.: Seven reasons why: A user's guide to transparency and reproducibility **13**(2), e0001 (2019). https://doi.org/10.1590/1981-3821201900020001, http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1981-38212019000200400&tlng=en
10. Freese, J., Peterson, D.: Replication in social science. Annual Review of Sociology **43**(1), 147–165 (2017). https://doi.org/10.1146/annurev-soc-060116-053450, https://doi.org/10.1146/annurev-soc-060116-053450
11. Goodman, S.N., Fanelli, D., Ioannidis, J.P.A.: What does research reproducibility mean? Science Translational Medicine **8**(341), 341ps12–341ps12 (2016). https://doi.org/10.1126/scitranslmed.aaf5027, https://www.science.org/doi/abs/10.1126/scitranslmed.aaf5027
12. Howison, J., Herbsleb, J.D.: Scientific software production: Incentives and collaboration. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. p. 513–522. CSCW '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1958824.1958904, https://doi.org/10.1145/1958824.1958904
13. Hoy, M.B.: Big data: An introduction for librarians. Medical Reference Services Quarterly **33**(3), 320–326 (2014). https://doi.org/10.1080/02763869.2014.925709, pMID: 25023020
14. Hussein, H., Oelen, A., Karras, O., Auer, S.: KGMM - a maturity model for scholarly knowledge graphs based on intertwined human-machine collaboration. In: Tseng, Y.H., Katsurai, M., Nguyen, H.N. (eds.) From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries. pp. 253–269. Springer International Publishing (2022)
15. Key, E.M.: How are we doing? data access and replication in political science. PS: Political Science Politics **49**(2), 268–272 (2016). https://doi.org/10.1017/S1049096516000184

16. Knublauch, H., Kontokostas, D.: Shapes constraint language (shacl). W3C Candidate Recommendation **11**(8) (2017)
17. Krawczyk, M., Reuben, E.: (un)available upon request: Field experiment on researchers' willingness to share supplementary materials **19**(3), 175–186 (2012). https://doi.org/10.1080/08989621.2012.678688, publisher: Taylor & Francis
18. Leek, J.T., Peng, R.D.: Reproducible research can still be wrong: Adopting a prevention approach. Proceedings of the National Academy of Sciences **112**(6), 1645–1646 (2015). https://doi.org/10.1073/pnas.1421412111, `https://www.pnas.org/doi/abs/10.1073/pnas.1421412111`
19. Lucas, J.W., Morrell, K., Posard, M.: Considerations on the 'replication problem' in sociology **44**(2), 217–232 (2013). https://doi.org/10.1007/s12108-013-9176-7, `https://doi.org/10.1007/s12108-013-9176-7`
20. Markowetz, F.: Five selfish reasons to work reproducibly **16**(1), 274 (2015). https://doi.org/10.1186/s13059-015-0850-7, `https://doi.org/10.1186/s13059-015-0850-7`
21. Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P.A.: A manifesto for reproducible science **1**, 0021 (2017). https://doi.org/10.1038/s41562-016-0021
22. Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D.P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E.L., Simonsohn, U., Soderberg, C., Spellman, B.A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.J., Wilson, R., Yarkoni, T.: Promoting an open research culture. Science **348**(6242), 1422–1425 (2015). https://doi.org/10.1126/science.aab2374, `https://www.science.org/doi/abs/10.1126/science.aab2374`
23. Peng, R.D.: Reproducible research in computational science **334**(6060), 1226–1227 (2011). https://doi.org/10.1126/science.1213847
24. Savage, C.J., Vickers, A.J.: Empirical study of data sharing by authors publishing in PLoS journals **4**(9) (2009). https://doi.org/10.1371/journal.pone.0007078, `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007078`
25. Stocker, M., Oelen, A., Jaradeh, M.Y., Haris, M., Oghli, O.A., Heidari, G., Hussein, H., Lorenz, A.L., Kabenamualu, S., Farfar, K.E., Prinz, M., Karras, O., D'Souza, J., Vogt, L., Auer, S.: FAIR scientific information with the open&nbsp;research knowledge graph **1**(1), 19–21 (2023). https://doi.org/10.3233/FC-221513, `https://content.iospress.com/articles/fair-connect/fc221513`
26. Vines, T.H., Albert, A.Y.K., Andrew, R.L., Débarre, F., Bock, D.G., Franklin, M.T., Gilbert, K.J., Moore, J.S., Renaut, S., Rennison, D.J.: The availability of research data declines rapidly with article age **24**(1), 94–97 (2014). https://doi.org/10.1016/j.cub.2013.11.014, `https://www.cell.com/current-biology/abstract/S0960-9822(13)01400-0`, publisher: Elsevier
27. 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., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 1–9 (2016). https://doi.org/10.1038/sdata.2016.18