MANCHESTER
1824
The University of Manchester

The University of Manchester Research

Are We Cobblers without Shoes? Making Computer Science Data FAIR

DOI:
10.1145/3528574

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Noy, N., & Goble, C. (2023). Are We Cobblers without Shoes? Making Computer Science Data FAIR. *Communications of the ACM*, *66*(1), 36-38. https://doi.org/10.1145/3528574

Published in:
Communications of the ACM

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:18. Jan. 2023

Are we cobblers without shoes? Making Computer Science data FAIR

NATASHA NOY, Google Research, USA
CAROLE GOBLE, University of Manchester, United Kingdom

ACM Reference Format:
Natasha Noy and Carole Goble. 2021. Are we cobblers without shoes? Making Computer Science data FAIR. 1, 1 (December 2021), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

We have recently asked a colleague to share a dataset that they published along with their paper at one of the ACM conferences. The paper had the "Artifacts available" badge[1] in the ACM Digital Library and the dataset and software were published, making the research in the paper reproducible. Yet, the instructions to get the dataset required several steps rather than just a link: log in, find the paper, click on a tab, scroll, get to the dataset. It was much better than receiving the data by email. Yet in many other research disciplines—biology, geophysics, biodiversity, social sciences, cultural heritage—open access to and sharing of data and other research artifacts are expected and streamlined. So how did Computer Science researchers get behind many other sciences in how we think about sharing data and other artifacts from our research?

Let's start by distinguishing three different aspects of data sharing: (1) open data, (2) data required for reproducibility of published research, and (3) data as a first-class citizen in scientific discourse. And while all three aspects are related, they are not the same: a dataset can be open but not citable or easily discoverable, for example. Or a dataset may be findable and interoperable, but not open.

Of the three aspects of data sharing that we mentioned, **open data**, or data that is available for free under appropriate licenses, is probably most familiar to many CS researchers: most of us are steeped in open-source software and understand and appreciate the value of sharing our research in an open way. Open data is just as important and is the bedrock of data-driven research and innovation as practiced by, for example, modern bioscience.[2]

Reproducibility in research is critical for trust and transparency [5]. ACM encourages[3] reproducibility of research through badges for papers that have data, code, or other artifacts available. Researchers in some fields within Computer Science were both instrumental in defining what reproducibility in computing means and in pushing their fields to embrace it. These fields include Databases[4,5], Machine Learning [6], Information Retrieval [6] where conferences have reproducibility tracks and where there is an expectation that research will be reproducible. Coincidentally (or maybe

[1] https://www.acm.org/publications/policies/artifact-review-badging
[2] https://elixir-europe.org/news/new-report-shows-open-data-heart-innovation
[3] https://www.acm.org/publications/policies/artifact-review-badging
[4] https://reproducibility.sigmod.org/
[5] https://vldb.org/pvldb/reproducibility/
[6] https://github.com/lintool/IR-Reproducibility

Authors' addresses: Natasha Noy, Google Research, USA, natashafn@acm.org; Carole Goble, University of Manchester, United Kingdom, carole.goble@manchester.ac.uk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
Manuscript submitted to ACM

not) these are fields where access to data for training, benchmarking and algorithm bake-offs is critical. Reproducibility usually entails data, code, and computational environment being accessible to readers of a paper. Note that reproducibility does not necessarily imply that the data is open or that it is citable or discoverable by itself, without the paper that it supplements. Indeed, finding or citing these types of datasets independent of the papers does not necessarily make sense in many cases: the datasets may not be useful outside of the context of reproducing the research in the paper.

Finally, thinking of data as a **first-class citizen** is the third aspect of sharing. Well-defined and well-described datasets, machine-learning models, and other artifacts become an engine for new papers and research; they can serve as a starting point for the next advance; they can inform new research questions and provide benchmarks to compare against. In other words, data, models, and software that we share as the result of our work should itself be a first-class citizen—and should be rewarded accordingly [3]. If we treat contributions of novel well-documented datasets and software packages with the same reverence that we treat papers, researchers will be more motivated to make these contributions. This goal is somewhat independent from the idea of reproducibility, though they are often conflated: in both cases we make data and software accessible. When we think about reproducibility, we think about validating the research that has been published. When we think of data and software as independent artifacts, we think about the ways that they can be reused for new research.

In many disciplines, the approach to data captured by the acronym FAIR has taken hold: data should be Findable, Accessible, Interoperable, and Reusable [8]. Making data FAIR elevates it to being first-class citizens in scientific discourse: datasets are valuable contributions by themselves, and others can reuse, cite, and evaluate them. FAIR data is complementary to the notion of reproducibility of research: data being FAIR is about such things as metadata, licensing, data being in a public persistent repository. Data being FAIR is also complementary to it being open: datasets published in an open repository with no metadata or license is not FAIR and does not allow proper reuse. At the same time, a dataset may have a license that defines constraints on its reuse, and still be FAIR.

In the last few years, many scientific communities have adopted the notion of FAIR data as the core of how they will share their research. For example, essentially all journals that publish papers in **geosciences** (which includes earth and planetary sciences, climate research, etc.) require [1] all authors to make all data that support the conclusions in their papers available in publicly accessible repositories that follow the FAIR principles.[7] These changes "elevate data to valuable research contributions rather than the files that are shoved in as an afterthought." [7] Major journals in fields such as Material Science and Biology, as well as almost all of the Nature journals.[8] Researchers in fields outside of Computer Science are often familiar with such platforms as Code Ocean,[9] which enable publication of research objects encapsulating data, software, and computational environment and making these objects citable. Government entities from OECD[10] and UNESCO[11] to national governments[12] have embraced the notion of FAIR data for any research data that is created with public funds.

How are we doing in Computer Science? The short answer is "not good." For example, of the 119 ACM conferences,[13] only **five**[14] encourage their authors to follow FAIR data principles and to submit data and software in public repositories that support these principles. That's less than 4%. Even for reproducibility, the situation is only slightly better: of the

[7] https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/
[8] https://www.springernature.com/gp/authors/research-data-policy/journal-policies-and-services
[9] https://codeocean.com/
[10] https://www.oecd.org/sti/enhanced-access-to-publicly-funded-data-for-science-technology-and-innovation-947717bc-en.htm
[11] https://en.unesco.org/science-sustainable-future/open-science
[12] https://www.inrae.fr/en/news/second-national-plan-open-science-inrae-manage-recherche-data-gouv-national-research-data-platform
[13] https://dl.acm.org/conferences
[14] The five conferences are: the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE) ; ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM); Automated Software Engineering

remaining 114 ACM conferences, only nineteen mention any sort of artifact submission in their calls for papers—and that's with ACM having an Artifact evaluation policy and support for it. 80% of the ACM conferences don't mention anything about sharing data. And while some of these are theory conferences where there are no research artifacts beyond the paper itself, the vast majority are not. Some of the non-ACM conferences such as NeurIPS[15] and ICML[16] also treat datasets and code associated with the papers, particular dataset papers, as first-class objects.

So, what would it mean in practice to have Computer Science venues require that research artifact submissions follow the FAIR principles?

Identifiers. Consider how often you have published data on your own web site or submitted a zip file along with your paper? Such datasets lack identifiers that are either persistent (a URL to your site will change) or dereferenceable (can we always find a dataset by its identifier?). The publishing industry has long since found a solution for persistent reference to artifacts: unique, persistent, dereferenceable identifiers. These identifiers provide three critical features: identifiers are unique, persistent, and dereferencable. We can refer to an artifact by a string of characters and numbers that uniquely identify it; there is a permanent URL that will always go to the main page of the artifact, even if that particular page moves somewhere. Digital object identifiers (DOIs), compact identifiers,[17] and similar schemes all serve this purpose.

Metadata, languages, and standards. Metadata is critical for both humans and tools to understand the data. Humans need to know how the data was created, who owns it, what are the constraints. Owners and providers provide an implicit authority signal. Machine-readable metadata makes the data discoverable. Standards such as schema.org and W3C DCAT allow this machine readable metadata to be embedded in the landing pages for datasets: the human-readable rendering of the page remains the same, whereas semantic metadata is embedded. This metadata may be as simple as the title and description of a dataset, or much more detailed, including spatial and temporal coverage, provenance, providers, etc. There are vocabularies developed by specific communities of practice that extend the metadata with the domain-specific terms. For instance, bioschemas,[18] by the life science community, or dataset metadata that the scientists in the Earth Science Information Partners (ESIP)[19] have agreed to. A recent survey provides a comprehensive analysis of metadata standards for computationally reproducible research [4]. A recent survey provides a comprehensive analysis of metadata standards for computationally reproducible research [4].

Licenses and access. Clear licenses make data and software reuse possible. However, a recent analysis of datasets on the Web found that 70% of datasets with machine-readable metadata come without an explicitly specified license [2]. And yet, in practice one cannot confidently reuse a dataset that does not have a license. Not having a license does not make a dataset "open": on the contrary, it prevents reuse by not giving others confidence of what they can and cannot do with a dataset. Creative Commons licenses [20] are a popular choice for datasets and there are a variety of choices for software. [21]

(ASE); the International Conference on Knowledge Capture (K-CAP); ACM Conference on Computer-supported cooperative work and Social Computing (CSCW)
[15] https://neurips.cc/Conferences/2021/PaperInformation/CodeSubmissionPolicy
[16] https://icml.cc/FAQ/authors-submit-data
[17] http://identifiers.org
[18] http://bioschemas.org
[19] https://www.esipfed.org
[20] https://creativecommons.org/licenses/
[21] https://www.software.ac.uk/resources/guides/choosing-open-source-licence

Repositories and permanence. . The final question is *where to publish?* The tendency among many CS researchers is to create our own Website, or to put it on our lab's page. However, these types of pages inevitably move (or so do people who own them). Anybody who wants to find a dataset mentioned in a reference several years later, may have trouble tracking it down. Thus, long-term availability is the first point to consider. Today, many dataset repositories (e.g., figshare, [22] Zenodo, [23] Data Dryad [24], Kaggle [25]) not only take care of providing long term access to the data, similar to how publishers do, but also have agreements with libraries [26] for preserving the data in perpetuity. Furthermore, these repositories make all other aspects of FAIR data sharing easier by generating metadata automatically. GitHub recently announced [27] the ability to cite their code repositories; repositories such as figshare, Zenodo, DataDryad, Kaggle, and others also enable embargoed and anonymized submissions while papers are being reviewed.

Will following *all* these guidelines make data FAIR? Not necessarily. A lot still depends on the social structures that we are yet to build around data publishing. How much is enough in terms of describing the conditions on how a dataset was created? How much do we need to know about the samples, how they were collected, how they were annotated? If a paper describes the creation of a dataset, should we be citing the paper or the dataset? How do we incorporate versioning and provenance of the data and code? Should the sharing and reproducibility be simply a "push of the button"? How can we create features in the repositories that add value to the data and code that we find there, for example, by suggesting related datasets, finding models that can be applied to a dataset that we found, giving nuanced and useful metrics on the level and types of reuse. All these issues are actively discussed and solutions proposed in CODATA, RDA, ReSA, AGU, Force11 and other fora where researchers who handle data and produce code gather. But not Computer Science.

As we hopefully move from just a handful of Computer Science conferences and journals requiring that their artifact submissions follow the open-science principles, to having this a standard practice in the community, perhaps conference and journals should have their own badges on how much they support or require publication of software and data and whether the requirements follow the FAIR principles. After all, Computer Science researchers are often the ones developing and publishing metadata standards, provenance frameworks, efficient data and code repository infrastructures. We can use these tools to make our own artifacts FAIR. As we make and mend the shoes for everybody else, we, as Computer Scientists, should wear our own shoes.

REFERENCES

[1] 2019. FAIR play in geoscience data. *Nature Geoscience* 12, 961 (2019). https://doi.org/10.1038/s41561-019-0506-4

[2] Omar Benjelloun, Shiyu Chen, and Natasha Noy. 2020. Google dataset search by the numbers. In *International Semantic Web Conference*. Springer, 667–682.

[3] Amanda Casari, Katie McLaughlin, Milo Z Trujillo, Jean-Gabriel Young, James P Bagrow, and Laurent Hébert-Dufresne. 2021. Open source ecosystems need equitable credit across contributions. *Nature Computational Science* 1, 1 (2021), 2–2.

[4] Jeremy Leipzig, Daniel Nüst, Charles Tapley Hoyt, Karthik Ram, and Jane Greenberg. 2021. The role of metadata in reproducible computational research. *Patterns* 2, 9 (2021), 100322. https://doi.org/10.1016/j.patter.2021.100322

[5] National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and replicability in science.* National Academies Press. https://doi.org/10.17226/25303

[6] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Hugo Larochelle. 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *CoRR* abs/2003.12206 (2020). arXiv:2003.12206 https://arxiv.org/abs/2003.12206

[22] https://figshare.com/
[23] https://zenodo.org/
[24] https://datadryad.org/
[25] https://www.kaggle.com/datasets
[26] https://help.figshare.com/article/preservation-and-continuity-of-access-policy
[27] https://twitter.com/natfriedman/status/1420122675813441540

[7] Shelley Stall, Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson, and Lesley Wyborn. 2019. Make scientific data FAIR. *Nature* 570 (2019), 27–29. https://doi.org/10.1038/d41586-019-01720-7

[8] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9. https://doi.org/10.1038/s43588-020-00011-w