1st Conference on Research Data Infrastructure

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.299

Published: 07 Sept. 2023

Open Science Best Practices in Data Science and Artificial Intelligence

Ekaterina Borisova[1] [https://orcid.org/0000-0002-3447-9860],
Raia Abu Ahmad[1] [https://orcid.org/0009-0004-8720-0116], and
Georg Rehm[1] [https://orcid.org/0000-0002-7800-1893]

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany

Abstract: In the past years, scientific research in Data Science and Artificial Intelligence has witnessed vast progress. The number of published papers and digital objects (e. g., data, code, models) is growing exponentially. However, not all research artefacts fulfill the criteria of being findable, accessible, interoperable and reusable (FAIR), contributing to a rather low level of reproducibility of experimental findings reported in scholarly publications and to the reproducibility crisis. In this paper, we focus on Data Science and Artificial Intelligence Open Science best practices, i. e., a set of recommendations that eventually contribute to the management and development of digital artefacts that are as FAIR as possible. While several guidelines exist, we add best practices for the FAIR collection, processing, storing and sharing of scholarly findings via Research Knowledge Graphs. The final list of recommendations will be available on the NFDI4DS website as an interactive web application.

Keywords: FAIR, Reproducible Research, Open Science

1 Introduction

The past years have seen rapid progress in the fields of Data Science (DS), Machine Learning (ML) and Artificial Intelligence (AI) with neural methods becoming the state-of-the-art in a wide range of research areas such as natural language processing and computer vision. Contemporary computational methods usually consist of code, ML models and data used for their training and evaluation. With an ever-increasing amount of newly appearing ML techniques and datasets, the question of how to make digital objects (especially code, data, models, software) *findable*, *accessible*, *interoperable* and *reusable* (FAIR) [1] to ensure the transparency and reproducibility of research results has never been more relevant and important than today [2], [3].

Recent studies demonstrated that scientific ML and AI pipelines often lack fine-grained documentation. Details on model hyperparameters, data pre-processing steps, evaluation metrics, dependencies, train/test splits, biases, annotation procedures, etc. are either only documented partially or not at all [4], [5]. The absence of this information complicates the validation, replication and improvement of previous findings, i. e., it significantly hinders scientific progress. Furthermore, it is quite common that code, software and (meta)data are missing or not specified or cited at all in scientific papers [6]. Yousuf et al. [6] show that only about 30% of Computer Science papers from arXiv include links to source code. Since academic literature and search engines are the main data discovery sources for researchers [7], links between digital objects and publications are crucial to guarantee accessibility. In addition, code, models or data are not always publicly available due to privacy, legal, ethical, commercial or copyright restrictions (e. g., medical data, Generative Pre-trained Transformer, GPT, [8] models developed by OpenAI [9], etc.). This factor also contributes to the challenge of FAIR research. All of the aforementioned issues contribute to the *reproducibility crisis* [2], [3] because it is getting increasingly difficult – in many cases it is already impossible – to reuse results and to reproduce state-of-the-art methods.

Reproducibility concerns gave rise to a series of workshops [10]–[15], checklists [16]–[19] and a handbook [20] on FAIR scientific research. Moreover, the availability of digital objects has become a common criterion for evaluating paper submissions at conferences (e. g., NAACL [21], ACL [22]) and in journals (e. g., Nature [23]). However, despite the proposed measures, digital artefacts still tend to be published with incomplete descriptions of their provenance, quality or dependencies [6], [24]. As an extreme example, OpenAI's technical report on GPT-4 [25] does not provide details on the model's architecture, dataset construction method or training procedure. This lack of information affects not only research replication but also leads to ethical concerns since it is impossible to verify the use/misuse of personal data during the model training (e. g., private messages). It is essential to encourage scholars and companies to use and develop open-source ML models (such as BLOOM [26]).

2 Open Science Best Practices in Data Science and Artificial Intelligence

To promote the idea of reproducible research, we propose *Open Science best practices especially geared towards Data Science and Artificial Intelligence research*, i. e., recommendations for ensuring a FAIR lifecycle of digital objects. The best practices were collected and summarised based on previous developments in research data and software management (e. g., Gebru et al. [27], Lamprecht et al. [28], Barker et al. [29], Pineau et al. [17], Rogers et al. [18], Dodge et al. [19], Rehm [30], NeurIPS 2021 Paper Checklist Guidelines [16], etc.). It is worth noting that our recommendations are developed primarily for scholars who plan to publish their research, we encourage credibility and findability in science. The core of our best practices constitutes topics ranging from the FAIR collection and processing of data to the distribution, validation and maintenance of (meta)data, code, models and software. For example, our recommendations for code, models and software distribution include but are not limited to:

- Make your code, models or software publicly available. Publish them in an appropriate, recognised and trusted [31] repository. We encourage the use of open source and open access repositories which guarantee the persistent identification (e. g., DOI, PID), long-term availability and authenticity protection of digital artefacts (e. g., Software Heritage [32]);
- Publish code, models or software with rich metadata using an appropriate metadata format;

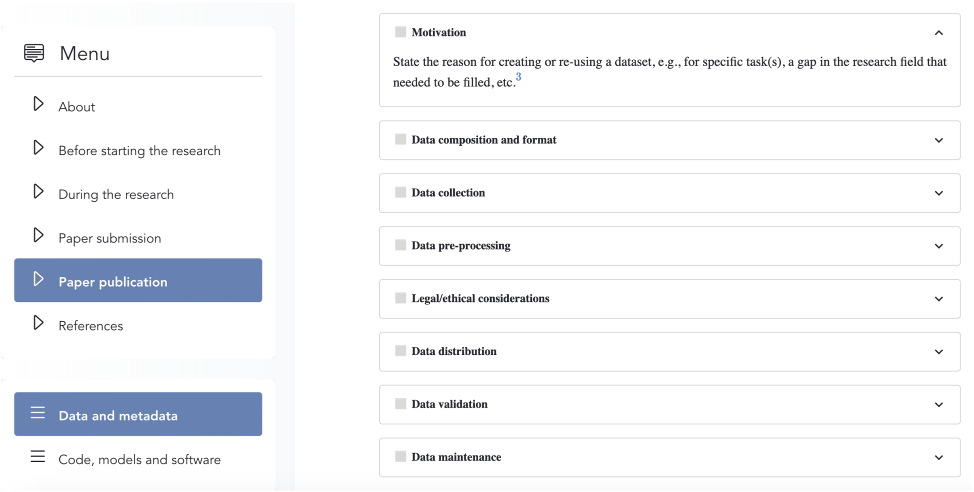
- Make sure the code can be run out of the box (time and machine independent). Make use of Docker [33] containers and eventually consider publishing them through a community platform such as European Language Grid (ELG) [34].

In addition, the proposed best practices have guidelines for transparent management and sharing of scientific artefacts via Research Knowledge Graphs (RKGs) [35] such as the Open Research Knowledge Graph (ORKG) [36] or the Semantic Scholar Academic Graph (S2AG) [37]. RKGs allow the representation of scientific results and contributions through structured, semantically rich, interlinked knowledge graphs. RKGs are aimed to establish an efficient search across scholarly findings so that researchers can gain an overview of recent developments and are able to compare their findings. In this sense, it has become increasingly crucial that scientific resources are stored, shared, harvested and processed in a FAIR way. For instance, we recommend researchers to label their contributions (e. g., research problem, objective, method, etc.) in the LaTeX file using the SciKGTeX [38] package to allow the automatic extraction and import of this metadata into RKGs. To the best of our knowledge, the topic of RKGs is not covered by existing resources such as The Turing Way handbook [20] or checklists for reproducible ML and AI research [16]–[19].

To improve user experience and user adoption, especially with regard to junior researchers, our recommendations are aligned with the typical research timeline associated with the development of scientific articles and split into the following four phases:

1. *Before starting the research*
2. *During the research*
3. *Paper submission*
4. *Paper publication*

The DS and AI Open Science best practices we collected are presented in the form of an interactive Streamlit application [39] (see Figure 1). The recommendations will be made available on the NFDI4DS website [40].



State the reason for creating or re-using a dataset, e.g., for specific task(s), a gap in the research field that needed to be filled, etc.[3]

Figure 1. The Open Science Best Practices in Data Science and AI Streamlit application

Data availability statement

No data is shared in the paper.

Author contributions

Georg Rehm conceived the original idea. Ekaterina Borisova and Raia Abu Ahmad performed the research activity planning and execution. Ekaterina Borisova, Raia Abu Ahmad and Georg Rehm wrote this paper.

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS) as part of the non-profit association National Research Data Infrastructure (NFDI e.V.). The NFDI is funded by the Federal Republic of Germany and its states. The paper received funding through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). The authors wish to thank both for funding and support. A special thanks goes to all institutions and actors engaging for the association and its goals.

References

[1] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, “The FAIR Guiding Principles for Scientific Data Management and Stewardship,” *Scientific Data*, vol. 3, no. 160018, Mar. 2016. DOI: 10.1038/sdata.2016.18.

[2] A. Belz, S. Agarwal, A. Shimorina, and E. Reiter, “A Systematic Review of Reproducibility Research in Natural Language Processing,” in *Proc. of the 16th Conf. of the Europ. Chap. of the Assoc. for Comput. Ling.*, Association for Computational Linguistics, May 2021, pp. 381–393. DOI: 10.18653/v1/2021.eacl-main.29.

[3] M. Hutson, “Artificial Intelligence Faces Reproducibility Crisis,” *Science*, vol. 359, no. 6377, pp. 725–726, Feb. 2018. DOI: 10.1126/science.359.6377.725.

[4] F. D. Maurizio, C. Paolo, and J. Dietmar, “Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches,” in *Proc. of the 13th Assoc. for Comput. Machinery (ACM) Conf. on Recomm. Systems*, Copenhagen, Denmark: Association for Computing Machinery, Sep. 2019, pp. 101–109. DOI: 10.1145/3298689.3347058.

[5] L. Rupprecht, J. C. Davis, C. Arnold, Y. Gur, and D. Bhagwat, “Improving Reproducibility of Data Science Pipelines Through Transparent Provenance Capture,” in *Proc. VLDB Endow.*, VLDB Endowment, Aug. 2020, pp. 3354–3368. DOI: 10.14778/3415478.3415556.

[6] R. B. Yousuf, S. Biswas, K. K. Kaushal, *et al.*, “Lessons from Deep Learning Applied to Scholarly Information Extraction: What Works, What Doesn’t, and Future Directions,” 2022. arXiv: 2207.04029 [cs.IR].

[7] K. Gregory, P. Groth, A. Scharnhorst, and S. Wyatt, “Lost or Found? Discovering Data Needed for Research,” *Harvard Data Science Review*, vol. 2, no. 2, May 2020. DOI: 10.1162/99608f92.e38165eb.

[8] A. Radford and K. Narasimhan, “Improving Language Understanding by Generative Pre-Training,” 2018.

[9] OpenAI. “OpenAI.” Accessed: 2023-04-09. (2023), [Online]. Available: https://openai.com.

[10] A. Lucic, M. Bleeker, S. Bhargav, *et al.*, "Towards Reproducible Machine Learning Research in Natural Language Processing," in *Proc. of the 60th Annual Meeting of the Assoc. for Comput. Ling.: Tutorial Abstracts*, Association for Computational Linguistics, Jun. 2022, pp. 7–11. DOI: 10.18653/v1/2022.acl-tutorials.2.

[11] GO FAIR. "M4M Workshop." Accessed: 2023-04-08. (2018), [Online]. Available: https://www.go-fair.org/events/m4m-workshop/.

[12] GO FAIR. "M4M #2: Preclinical trials + M4M #3: Funders." Accessed: 2023-04-08. (2019), [Online]. Available: https://www.go-fair.org/events/m4m-2-preclinical-trials-m4m-3-funders/.

[13] GO FAIR. "The Second GO FAIR Workshop for the German Research Community." Accessed: 2023-04-08. (2018), [Online]. Available: https://www.go-fair.org/2018/10/08/on-the-road-to-fair/.

[14] GO FAIR. "The 3rd Germany GOes FAIR Workshop for the German Research Community." Accessed: 2023-04-08. (2019), [Online]. Available: https://www.go-fair.org/2019/06/04/3rd-germany-goes-fair-workshop-report/.

[15] OpenAIRE and RDA Europe and FAIRsFAIR and EOSC-hub. "Services to Support FAIR Data." Accessed: 2023-04-08. (2019), [Online]. Available: https://eosc-portal.eu/events/workshop-series-services-support-fair-data.

[16] NeurIPS. "NeurIPS 2021 Paper Checklist Guidelines." Accessed: 2023-04-08. (2021), [Online]. Available: https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist.

[17] J. Pineau, P. Vincent-Lamarre, K. Sinha, *et al.*, "Improving Reproducibility in Machine Learning Research (A Report from the Neurips 2019 Reproducibility Program)," *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, Jan. 2021.

[18] A. Rogers, T. Baldwin, and K. Leins, "'Just What Do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP," in *Findings of the Assoc. for Comput. Ling.: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4821–4833. DOI: 10.18653/v1/2021.findings-emnlp.414.

[19] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, "Show Your Work: Improved Reporting of Experimental Results," in *Proc. of the 2019 Conf. on Empirical Methods in NLP and the 9th Intern. Joint Conf. on NLP (EMNLP-IJCNLP)*, Association for Computational Linguistics, Nov. 2019, pp. 2185–2194. DOI: 10.18653/v1/D19-1224.

[20] The Turing Way Community. "The Turing Way: A Handbook for Reproducible, Ethical and Collaborative Research (1.0.2)." Accessed: 2023-04-08. (2021), [Online]. Available: https://doi.org/10.5281/zenodo.7625728.

[21] NAACL. "NAACL 2021 Reproducibility Checklist." Accessed: 2023-04-08. (2021), [Online]. Available: https://2021.naacl.org/calls/reproducibility-checklist/.

[22] ACL. "ARR Responsible NLP Research Checklist." Accessed: 2023-04-08. (2021), [Online]. Available: https://aclrollingreview.org/responsibleNLPresearch/.

[23] Nature. "Reporting Standards and Availability of Data, Materials, Code and Protocols." Accessed: 2023-04-08. (2023), [Online]. Available: https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards.

[24] A. Spirling, "Why Open-source Generative AI Models are an Ethical Way Forward for Science," *Nature*, vol. 616, no. 413, Apr. 2023. DOI: 10.1038/d41586-023-01295-4.

[25] OpenAI, "GPT-4 Technical Report," 2023. arXiv: 2303.08774 [cs.CL].

[26] T. L. Scao, A. Fan, C. Akiki, *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," 2023. arXiv: 2211.05100 [cs.CL].

[27] T. Gebru, J. Morgenstern, B. Vecchione, *et al.*, "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021. DOI: 10.1145/3458723.

[28] A. L. Lamprecht, L. Garcia, M. Kuzak, *et al.*, "Towards FAIR Principles for Research Software," *Data Science*, vol. 3, no. 1, pp. 37–59, Jun. 2020. DOI: 10.3233/DS-190026.

[29] M. Barker, N. P. Chue Hong, D. S. Katz, *et al.*, "Introducing the FAIR Principles for Research Software," *Scientific Data*, vol. 9, no. 622, Oct. 2022. DOI: 10.1038/s41597-022-01710-x.

[30] G. Rehm, "The Language Resource Life Cycle: Towards a Generic Model for Creating, Maintaining, Using and Distributing Language Resources," in *Proc. of the 10th Intern. Conf. on Lang. Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), May 2016, pp. 2450–2454.

[31] D. Lin, J. Crabtree, I. Dillo, *et al.*, "The TRUST Principles for Digital Repositories," *Scientific Data*, vol. 7, Dec. 2020. DOI: 10.1038/s41597-020-0486-7.

[32] Software Heritage. "Software Heritage." Accessed: 2023-07-06. (2023), [Online]. Available: https://zenodo.org.

[33] Docker. "Docker." Accessed: 2023-04-14. (2023), [Online]. Available: https://www.docker.com.

[34] ELG. "European Language Grid." Accessed: 2023-04-14. (2023), [Online]. Available: https://live.european-language-grid.eu.

[35] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, and M. E. Vidal, "Towards a Knowledge Graph for Science," in *Proc. of the 8th Intern. Conf. on Web Intelligence, Mining and Semantics*, ser. WIMS '18, Novi Sad, Serbia: Association for Computing Machinery, Jun. 2018, pp. 1–6. DOI: 10.1145/3227609.3227689.

[36] M. Y. Jaradeh, A. Oelen, K. E. Farfar, *et al.*, "Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge," in *Proc. of the 10th Inter. Conf. on Knowledge Capture*, Marina Del Rey, CA, USA: Association for Computing Machinery, Sep. 2019, pp. 243–246. DOI: 10.1145/3360901.3364435.

[37] R. M. Kinney, C. Anastasiades, R. Authur, *et al.*, "The Semantic Scholar Open Data Platform," 2023. arXiv: 2301.10140 [cs.DL].

[38] C. Bless, I. Baimuratov, and O. Karras, "SciKGTeX – A LaTeX Package to Semantically Annotate Contributions in Scientific Publications," 2023. arXiv: 2304.05327 [cs.DL].

[39] Streamlit. "Streamlit." Accessed: 2023-04-13. (2023), [Online]. Available: https://streamlit.io.

[40] NFDI4DS. "NFDI4DataScience." Accessed: 2023-04-12. (2023), [Online]. Available: https://www.nfdi4datascience.de.