ing.grid

DATA MANAGEMENT LETTER

Critically thinking about the reusability of (meta)data

Izadora Silva Pimenta [1]

1. Chair of Fluid Systems, Technische Universität Darmstadt, Darmstadt.

Date Submitted:
2024-03-06

Date Published:
2024-03-18

DOI:
doi.org/10.48694/inggrid.3945

Keywords:
Inggrid, Critical Data Studies, Data Literacy, Data Ethics, Data Reusability, RDM

Corresponding Author:
Izadora Silva Pimenta
izadora.pimenta@tu-darmstadt.de

Biography. I have a PhD in Digital Linguistics (TU Darmstadt). I hold an MA in Applied Linguistics (University of Campinas) and a bachelor's in Journalism (PUC-Campinas). Along my research path, I have been strongly connected to Systemic-Functional Linguistics, Appraisal Theory and Corpus Linguistics studies. Currently, I work at TU Darmstadt, as a Research Associate (Chair of Fluid Systems), and at the Gender Consulting for Research Networks (Gender Equality Office). I am a Managing Editor for ing.grid.

How much consideration are we giving to the (Meta)data we produce? Naturally, the FAIR principles lead us through several steps in which we can elevate data management to a scientific effort in its own right, as ing.grid shares and advocates. However, reflecting on what the (meta)data does and does not encompass is also a commendable endeavour. Citing the researcher Joy Buolamwini in the documentary "Coded Bias", data is destiny [1]. Data is a relationship we can make and put to use [2]. Taking responsibility for this (meta)data and striving to make it as transparent as possible is also a crucial step towards ensuring its reusability.

Engaging in critical thinking and taking ownership of the (meta)data we generate and disseminate not only enhances its worth but also steers us towards innovative pathways. Several approaches, such as the CARE Principles [3] [4] and the Feminist Data Manifesto [2], make us think of data as a resource to be cared for and cultivated [2], going beyond the colonial extraction logic. To achieve this, we must consider the narrative of our (meta)data, the stakeholders involved in its generation, and the societal values embedded within it. Who is generating the data? For whom is it intended? Do we contemplate the ramifications of this data? Is our focus solely on data generation without fully realising its potential?

Sarah Ciston, author of a guide on managing machine learning datasets, acknowledges that datasets that encompass diverse perspectives —- meaning, where feasible, those datasets that incorporate interdisciplinary and intersectional[1] communities in "designing, developing, implementing, and evaluating your work" [5] —- can offer a more robust approach to working with your data. Furthermore, they recognise that critical practices are becoming standard in many conferences and journals. Understanding that datasets can never be neutral ("taking no position on a dataset's ethical question is still taking a position", as Ciston reminds us), it is imperative to

1. Intersectionality is a term coined by Kimberle Crenshaw in 1989. It has to be of how interlocking systems of power affect those who are marginalised in society. To read more: https://www.law.columbia.edu/news/archive/kimberle-crenshaw-intersectionality-more-two-decades-later

bear certain considerations in mind:

> While it may be impossible to escape classification's worldviews entirely, with awareness of the underlying assumptions of classification and its impact on your processes, it becomes easier to make critical decisions that account for these contexts. [5, paragraphs 1466-1469]

(Meta)data is more than just a resource; it is a representation of specific contexts from our world. All processes carry social implications. Do we possess accurate data regarding the safety of seatbelts if we fail to consider all body types during our research [6]?[2] Without such considerations, and if we share our (meta)data without critical thinking, we also pave the way for inaccurate reproducibility. To optimise the reuse of (meta)data, as demanded by the FAIR principles, it is pertinent that we extend our thinking beyond mere management requirements. Generating data also entails taking responsibility for initiating their life cycle.

So, what is your dataset there for [5]? I like to ponder, having immersed myself in bell hooks' approach[3] to considering the care we show towards others, that when we talk about community, we are discussing an undeniable commitment and responsibility. We must nurture this community around us. Building a community around a subject is also tied to that notion. If we are contemplating novel ways of reshaping scholarly publications, we are also contemplating the knowledge we must share and learn from others. Advocating for transparency also involves considering the implications of this (meta)data.

When formulating the author guidelines for ing.grid, we considered some of these aspects. Describing the (meta)data's usability for the community and clarifying whether the data is sensitive to certain segments of our society are among the points already encompassed in our guidelines [7]. However, establishing a community for Research Data Management in Engineering Sciences can extend beyond these aspects. As we endeavour to place (meta)data at the forefront, we, as a community, hold the power to shape this trajectory.

As someone coming from the field of Systemic-Functional Linguistics, I am constantly reminded of J.R. Firth's maxim: "We shall know a word by the company it keeps" [8]. Reflecting on our communication processes through the lens of language, and viewing language as the mechanism for constructing meaning [9], every act of communication embodies a dynamic of power. (Meta)data represents power. By generating and disseminating it, we bear responsibility for the entire spectrum of communication associated with it. When we share our (meta)data within a community, we undertake the obligation to provide something that our community can either trust or challenge – thus, collectively advancing our understanding.

2. Some authors that discuss this issue further are Caroline Criado Perez in "Invisible Women: Data Bias in a World Designed for Men" and Rebekka Endler, in "Das Patriarchat der Dinge" (in German)
3. bell hooks (1952-2021) was an American author, theorist, educator and social critic working mainly in writings on race, feminism, class and education. Her name is always written without capital letters.

Conflict of interest

Izadora Silva Pimenta is a managing editor for ing.grid. This Data Management Letter does not necessarily reflect the opinion of ing.grid.

References

[1] S. Kantayya, *Coded bias*, 2020.

[2] "Feminist Data Manifest-No." (n.d.), [Online]. Available: https://www.manifestno.com/home (visited on 07/03/2024).

[3] "The CARE principles for indigenous data governance." (n.d.), [Online]. Available: https://www.gida-global.org/care (visited on 07/03/2024).

[4] S. R. Carroll, I. Garba, O. L. Figueroa-Rodríguez, *et al.*, "The CARE principles for indigenous data governance," *Data Science Journal*, vol. 19, pp. 43–43, 2020.

[5] S. Ciston, "A critical field guide for working with machine learning datasets," *Knowing Machines project*, K. C. Mike Ananny, Ed., 2023. [Online]. Available: https://knowingmachines.org/critical-field-guide (visited on 07/03/2024).

[6] S. Samuel. "Women suffer needless pain because almost everything is designed for men." (2019), [Online]. Available: https://www.vox.com/future-perfect/2019/4/17/18308466/invisible-women-pain-gender-data-gap-caroline-criado-perez (visited on 07/03/2024).

[7] "ing.grid author guidelines." (n.d.), [Online]. Available: https://www.inggrid.org/site/authorguidelines/ (visited on 07/03/2024).

[8] J. Firth, *Studies in Linguistic Analysis: Special Volume of the Philosogical Society* (Special Volume of the Philological Society). Blackwell, 1957.

[9] M. A. K. Halliday and R. Hasan, "Language, context, and text: Aspects of language in a social-semiotic perspective," 1989.