

the global infrastructuring of research data and its risks for (particularly qualitative and critical) education scholarship

Sigrid Hartong

This essay problematizes the so-far absence of broader critical debate on the governing dynamics around research data infrastructuring (not only) in education scholarship. More specifically, it refers to the growing role of large-scale data infrastructures as well as complexes of Research Data Management (RDM) that seek to foster ‘open’, reusable, and increasingly machine-readable (education) science. While research data infrastructuring in general, and RDM in particular, have so far been predominantly discussed from solutionist-positivist perspectives, this essay discusses the underestimated risks that lie in the governance dynamics of research data infrastructuring, including: (1) the gradual narrowing of what kind of education, but also what kind of education research become visible in and through data infrastructures (namely those that build on already dominant hierarchies in the field); (2) a gradual shift from doing research towards ‘doing data’, including a shift of political attention and research funding, as well as (3) effects of normalization, also triggered through extensive dataveillance, shareveillance, and ‘ethical monitoring’ of researchers. This particularly endangers genuinely qualitative and critical research in education.

Keywords: data infrastructures, infrastructuring, open science, politics, qualitative and critical education research, research data management (RDM)

Introduction: The Dream of ‘Open’ (Education) Research in an Age of Big Data

In 2018, the European Commission launched the *European Open Science Cloud* (EOSC) as part of the *European Data Science Policy* (Burgelman, 2021, p. 14). It was the first initiative of European scope that aimed to establish an integrated European, cloud-based data infrastructure to store, distribute, and make (re-)usable scientific knowledge (Giannoutakis & Tzovaras, 2017). While promoters of the initiative have been praising the cloud as a unique chance to connect cultures and practices of research (Veršić & Ausserhofer, 2019, p. 385), others emphasized the economic rationale behind the initiative, namely the need to create a European science industry – not only to serve the public but also and in particular businesses – that can compete with knowledge search engines such as Google or Alibaba (Burgelman, 2021, p. 7; Giannoutakis & Tzovaras, 2017, p. 220).

The European Union is by far not the only context where the digital infrastructuring of research data has become a prime focus of investment. Also, there has been a worldwide emergence or expansion of national research data strategies and data centers (Pryor, 2012). In 2013, the *US Office of Science and Technology Policy Memorandum* required open access for research data and products for many publicly funded institutes (Rafiq & Ameen, 2022, p. 392); similar approaches can be found in Canada or Australia. In Germany, national efforts to infrastructure education research data

in particular have already gained momentum after 2003, when the *Federal Department of Education and Research* (Bundesministerium für Bildung und Forschung, BMBF) started substantial investments in ‘evidence-based education policy’ (see also Jornitz, 2008, p. 206) and, related to that, in the more efficient storage and (re)usage of data gained through federal research funding schemes (see BMBF, 2023; Hartong et al., 2021). While the focus was hereby initially put on quantitative (e.g., survey) data, more calls emerged over the past years to also better infrastructure qualitative education research, as visible in the common recommendations for archiving, supplying, and reusing of education research data that was published by Germany’s three major education science associations in 2020.¹

This essay starts from the observation that such policies and programs, just as the discourse on research data infrastructuring in general, have so far been characterized by predominantly positive and simultaneously positivist perspectives. A first example is the common reference to *transparency* (e.g., Krammer & Svecnik, 2020, p. 267); that is, research data infrastructures enabling a more extensive evaluation of the validity and robustness of published research, because data (interpretations) can be tracked more easily. A second common argument is broader *accessibility* to data sources because data would ideally be stored centrally, can be easily downloaded, etc. (see also Gerecht & Kminek, 2018; Renbarger et al., 2023). Further arguments include the increase of *cost-efficiency* because research data can be used multiple times, or the *relief* for those who are frequently being researched (for instance, schools). An additional line of argumentation has emerged more recently: the seemingly limitless potential of *machine-based data analysis*. Big data techniques are described here as new ways of framing research questions, designing studies, and analyzing and visualizing data (Daniel, 2019, p. 101), which for disciplines such as education science would mean actively embracing data science methods and technologies (Daniel, 2019, p. 102). As Wilkinson and colleagues (2016) stated in their famous FAIR principles (FAIR meaning Findable, Accessible, Interoperable, and Reusable data), the idea is “[...] that computers must be capable of accessing a data publication autonomously, unaided by their human operators” (Mons et al., 2017, p. 51).

But also, for more human-centered ‘open science’ programs, preparing and processing of research data for machine-readability can be regarded as the core aspect of digital infrastructuring. The keyword here is *Research Data Management* (RDM) – also described as “data nurturing” (Heidorn, 2011, p. 669), “data curation” (Poole, 2015), or “data stewardship” (Mons et al., 2017, p. 51; see also Lewis & Hartong, 2022) –, a field of activity that stretches from data preparation, documentation, and formatting, metadata curation, to processing and transfer. Indeed, more and more guidelines, regulations, and RDM support structures for researchers have been installed in the field of education science over the past years, framed by extensive debates on how to solve organizational-practical questions or – as oftentimes in the case of education science – ethical challenges (for more detailed discussions see below).

While calls for more open, transparent, or accessible science may, as such, undoubtedly be legitimate, I argue here that the so-far predominantly solutionist-positivist understanding of ‘open science’ in general, and RDM in particular, substantially underestimates the actual risks that lie in the governance dynamics of research data infrastructuring (see also Hartong et al., 2021). Such risks include, amongst others, (1) the gradual narrowing of what kind of education, but also what kind of education research becomes visible in and through these infrastructures (namely those that build on already dominant hierarchies in the field); (2) a gradual shift from doing research towards ‘doing data’, including a shift of political attention and research funding, as well as (3) effects of

normalizing research practices, also triggered through extensive dataveillance, shareveillance, and ‘ethical monitoring’ of researchers. This particularly endangers genuinely qualitative and critical research in education.

Interestingly, such risks are nothing new but have similarly been discussed in other areas of (not only) educational datafication (e.g., performance monitoring), where the belief in large-scale data infrastructuring has caused a number of unintended consequences and harmful effects. Hence, this essay calls for a much broader engagement with such studies also in the context of ‘open science’ and RDM debates.

Why ‘Open Science’ is ‘Closed Science’, or: How Data Infrastructures Make Particular Worlds of Education (Research)

In his famous paper “Knowledge infrastructures and the inscrutability of openness in education” (2015), Richard Edwards problematized the oftentimes proclaimed open character of (here) data infrastructures. More specifically, he argued that infrastructuring technologies always “[...] work on the basis of ontologies, code, algorithms, and standards [...] that are not always open or opening” (Edwards, 2015, p. 251) but rather always install simultaneous opening and closing effects. Put differently; every openness carries closed-ness and vice versa (Edwards, 2015, p. 253).

Over the past years, there has been a large amount of research, including in the field of education, that critically scrutinized such closing effects in and through digital technologies, data infrastructures, and platforms (e.g., Decuyper et al., 2021; Kerssens & van Dijck, 2021). As this research has convincingly shown, APIs (*Application Programming Interfaces*) in platforms, for instance, strongly determine who can access data in which forms and under which conditions (see also Plantin et al., 2018, p. 10), thus closing off (other ways of) participation for many. Equally, datasets always “[...] emerge from and are enmeshed in power-laden semiotic systems” (Poirier, 2021, p. 1). In doing so, they always create particular educational worlds and exclude others, be it in the case of learning platforms or school management and monitoring systems (Hartong & Förschler, 2019). Small (and oftentimes hidden) pre-decisions can hereby have a large impact (Diesner, 2015), for instance, when defining rubrics of data, “allowed” forms of data input, visualization techniques (e.g., dashboarding), or inscribed pedagogies (Sefton-Greene & Pangrazio, 2021).

In the case of educational RDM and the platformization of research data, such world-making would, then, create a double layer in the sense that particular educational *and* particular research worlds become co-enacted together. As an example, predefinitions of how existent research is transferred into a centralized database (including the definition of machine-readable metadata), which rubrics define ‘searchability’, or how search outputs are visually displayed, have an immediate effect not only on which research is mostly found, but also on the viewpoints it offers on the education field. But also, the idea of the “research lifecycle” (Giannoutakis & Tzouvaras, 2017, p. 215), which can be frequently found in RDM documents, is a good example to illustrate how particular semiotic systems become inscribed into data infrastructures, namely understanding research as “[...] serial, uni-directional and occurring in a somewhat closed system” (Cox et al., 2018, p. 142). Grey and colleagues (2005) suggested the term “thinking and sorting devices” already a long time ago to describe such mechanisms, which equally form the basis for the governing dynamics of data infrastructures.

The Governing Dynamics of Data Infrastructures

Platforms' or databases' governmental logic aims at the regulation, standardization, and alignment of data practices. This may include obligatory data input, data formatting prescriptions, temporal regimes of data processing, or data documentation requests. Because such regulations are so crucial to the functioning of data infrastructures, but equally to the stabilization of the aforementioned fabricated worlds, they require enormous attention, which intensifies with every new challenge the infrastructure is facing (e.g., new data formats that need to be integrated). As Lewis and Hartong (2022) have shown, using the US EdFacts education monitoring infrastructure as an example, the result of that gradual intensification is a shift towards so-called second-order discourses and practices: the 'doing data' (or 'data stewardship', see above), which can become even more important than what is actually happening educationally. For instance, they show how, in EdFacts, professionals are confronted with a strict temporal regime of data submission,

[...] which forces them to constantly keep track of ongoing due dates, data submission preparation and reporting requirements. Simultaneously, these due dates and data collection cycles do not follow the regular school year, but instead employ a different logic of time pressure that continuously accelerates as data are requested from the states. [...] [S]tates that do not submit complete and timely data or data submission plans (which is meta-data on the state's data submission strategy) may be omitted from programmatic reports prepared for Congress, be cited for data submission failure or have federal funding punitively withheld. (Lewis & Hartong, 2022, p. 954).

It is such constantly growing amounts of data practice regulations and instruments that also back up *governance by data management* (Pels et al., 2018) in the field of 'open science'. One of many examples is data management plans, which McLeod and colleagues (2013, p. 84; cited in Poole, 2015, p. 110) defined as constituting "[...] the bedrock of a fit-for-purpose risk-managed approach to managing research data that is appropriate to its nature and potential value and meets any regulatory or other requirements". But also in the more specific field of education research infrastructures, a growing number of RDM regulations, training, templates, and support structures have emerged (see for the German case, for instance, <https://www.forschungsdaten-bildung.de>) that appear particularly 'thickening' where research practices are rather hard to fit into the infrastructures. An example of such thickening is context knowledge, which in qualitative, ethnographically oriented education research forms a crucial component of researchers' meaning-making but is undoubtedly hard to translate into standardized data formats. Still, a growing number of requests and guidelines try to tackle how contextualization could be formatted in more standardized ways and which "rich meta-data" (Bowers & Choi, 2023, p. 3) should be included in each dataset.

Another prominent area of struggle between infrastructuring and qualitative education research is ethical aspects: while data infrastructures need data to flow, particularly ethnographic education research typically builds on trust relationships and sensitive knowledge – oftentimes interactions with minors in vulnerable environments –, which requires strict data protection (see also below). But also here, various strategies are being extensively discussed to overcome this contradiction: anonymization, adapted consent forms, or ways to record classrooms that already have potential re-users of the data in mind.

Newer literature reveals an even better picture of where the field of RDM is currently heading, that

is, which instruments and strategies are being discussed to further enhance research data infrastructuring: while Krammer and Svecnik (2020), for instance, suggest so-called open science badges in order to further incentivize data inputs to centralized systems, Renbarger and colleagues (2023) present guidelines for peer reviewers to more strongly include RDM issues in their reviewing schemes. Strickert (2021) suggests installing an autonomous system of data peer review, and Robinson-García and colleagues (2016) even discuss the implementation of a data citation index which can serve as another proxy metric to measure research impact, similar to Web of Science (Robinson-García et al., 2016, p. 1350).

The Empowerment of New RDM Intermediaries...Human and Non-Human

As outlined so far, RDM is challenging, with each step requiring technical capacities, ethical considerations, legal issues, and guiding frameworks (Cox & Pinfield, 2014, pp. 300-302; Poole, 2015). As a consequence, numerous new actors have emerged and gained power within different contexts of data infrastructuring, actors whose job is primarily centered around (the enhancement of) data stewardship and mediation (see also Hartong, 2016).

Coming back to the initial example of the *European Open Science Cloud* (EOSC), the initiative equally included the founding of the *EOSC Association* (eo.sc.eu), which since 2021 has installed a gigantic network of actors, 13 task forces, and complex governance structures (all of them focusing on the implementation of the EOSC). In the aforementioned study of EdFacts, Lewis and Hartong (2022) equally identified various new roles and professions (e.g., Data Stewards, Data Coordinators, or Data Submitters) whose task is to protect the system by identifying and fixing data flaws, but also to motivate other participants of the infrastructure to care (better) for their data.

In the literature on ‘open’ science, it is oftentimes new departments in libraries and/or so-called data curation centers that are identified as central data mediators (Cox & Pinfield, 2014, p. 301; Poole, 2015, p. 101), together with new professions such as “Data Research and Information Visualization Specialists” (Uzwyshyn, 2022, p. 12). As Corral and colleagues (2012, p. 1) point out, in an era of digital infrastructuring, libraries need to provide services of digital curation in order to support researchers in meeting various requirements related to setting up data management plans, formatting files correctly, meeting meta-data standards, and so on (Corral, 2012, p. 10). At the same time, through such services, data mediators strongly co-shape and stabilize the infrastructuring process, including techniques of (soft) normalization in co-junction with the platforms in use (for instance, *DMP Online* as a service platform to support data management plans, see emponline.dcc.ac.uk). Poole (2015, p. 101) accordingly described the work of data curation centers as “adding value to digital data assets”, which, however, requires such centers to become involved early on in the original data production process (Poole, 2015, pp. 113–114).

At the same time, it seems important to acknowledge the very demanding (type of) labor of such intermediaries who are basically ruled by the same governance logic as the actors whose practices they seek to align. Plantin’s (2021) study on data archives as factories provides very informative insights here. In his ethnographic case study, he focuses on human data processors whose role is to work along the pipeline of data archiving between deposition and publication (Plantin, 2021, p. 4), including data cleaning/removing of data flaws, formatting according to templates, etc. Indeed, the study reveals a range of tensions in between which the data processors need to navigate, including frequent moments of alienation (e.g., when being reduced to box-ticking), onerous speed pressure, and the modulation of time, but also moments where active resistance is possible, and actively

exercised. Put differently; here we see again how governance by data management manifests as a complex assemblage in which (hidden) labor becomes increasingly relevant, just as ambiguous data practices that happen in such labor contexts.

Finally, also in the context of such transforming intermediary structures, the impact of AI is growing. There is a whole area of so-called AI library services, that is, the usage of AI tools in order to further automate practices such as research data sorting, data cleaning, metadata generation, or database searching (see for instance Uzwyshyn, 2022, who inter alia discusses the *Texas Data Repository* <https://dataverse.tdl.org>, as well as the platform www.openrefine.org). As such examples show, RDM practices themselves are becoming increasingly automated; that is, they transform under the new influence of, again, AI-related world-making procedures (Crawford, 2021). In other words, because each AI system, just like platforms or data infrastructures, more generally carries with it powerful selections of what is made visible and/or actionable as well as who is participating, we can assume that the RDM sphere will become gradually more infiltrated with such selections. Selections that are widely opaque not only to users but also increasingly to those who bring the technologies in.

Why Research Infrastructuring Poses Substantial Risks for (Particularly Qualitative and Critical) Education Scholarship

The fact that the discourse around ‘open science’ and RDM has, to a large extent, been driven by positive and positivist perspectives does not mean that one cannot also find more critical voices, particularly in the field of education research. For instance, Reh and colleagues (2020, p. 11) argue that understanding education research data as something to be managed falls dramatically short of acknowledging the actual methodological and methodical work substantiating data. Similarly, Pels and colleagues (2018, p. 396) problematize the data positivism in RDM more broadly, which entails a deceptive idea of data that strongly contradicts the very different meaning of research material to different people (see also Huber, 2019, p. 5). Equally, the idea of co-production of (here) ethnographic data brings, in their view, the ethical duty to scholars to control what goes public (Pels et al., 2018, p. 395).

While such concerns are undoubtedly extremely useful and important, many of them end with rather global suggestions, pointing, for instance, to still many open questions to be discussed (Reh et al., 2020), to the need for “extensive care and rigor” in the system (Poole, 2015, p. 115), to the preservation of freedom of choices for researchers to decide upon the sharing/re-usage of their data, or to the need to keep research funding allocation independent from such decisions (Reh et al., 2020, p. 11). In doing so, however, they tend to substantially underestimate the actual power of the governance dynamics that are *already* in place, which are *already* continuously expanding and thickening, and which are *already* showing a number of problematic effects. At least two of such effects seem particularly dramatic for education research and, within that field, even more so for qualitative and genuinely critical education research:

(1) Research infrastructures build on existing hegemonies and inequalities, reproduce them, and make them gradually less contestable.

As Crawford (2021) has extensively outlined in her book “The Atlas of AI”, data infrastructures commonly build on existent socio-political and economic hierarchies and hegemonies because they build on capital. In the field of education, numerous studies have further shown how the education

worlds that are being designed into platforms and data infrastructures today, while claiming to be diverse and inclusive, oftentimes enforce a narrowed, western-neoliberal, and engineering-oriented perspective (Williamson, 2018; Macgilchrist et al., 2023). Equally in the field of RDM, the large-scale programs are so far driven by particular actors with particular rationales and interests, embedded within politically and economically highly pre-structured contexts (see, for instance, the initial EOSC example, which strongly follows the idea of a competitive science industry). Also, data curation centers commonly emerge where there are already large amounts of funding or politically acknowledged expertise, just as standards are typically derived from already existing other standards. In the education science context, there has been ongoing critical debate on imbalances and hegemonies between different types of education research, which strongly affects how education is, in the end, made visible and governed. In that regard, worldwide discussions have particularly emerged around the growing dominance of ‘what works’-oriented, test-based, or quantitative education research, which has not only marginalized – scientifically, politically, and economically – education research that is more qualitative, theoretical, and critical in nature, but equally revealed massive normalizing effects on (young) researchers over the past decades (Gewirtz & Cribb, 2020). There is good reason to expect that with the gradual expansion and thickening of RDM governance, such asymmetries will become further enhanced and more and more difficult to challenge, given that young scholars depend on them for successful evaluation and given that it is exactly the aforementioned types of marginalized research which already fit least into the infrastructures.

(2) Research infrastructures install a constantly thickening system of dataveillance and shareveillance on researchers, which equally becomes more and more intertwined with ethical accountability.

As shown in the many examples above, research data infrastructuring indicates a quite substantial transformation of trust and, related to that, of transparency issues: while researchers are systematically distrusted to conduct robust research, to secure research quality through peer review (Hirschauer, 2014), or to communicate data (e.g., publications alone being rendered insufficient), technologies such as automated data integration or AI are trusted to fix such problems. In turn, while most large-scale databases or technologies such as AI commonly operate in ways that are at least partly opaque to most researchers, it is the same researchers that are, under the claim of transparency, increasingly being tracked in every single of their data practices. Or, as Burgelman (2021, p. 7) described it, “all science will become data-driven science and all [...] [scientific activity] will become a digital point or have a digital trace”. Already today, evidence shows that researchers tend to adjust and manipulate their own research when being aware of (future) surveillance (Imeri, 2017; Strickert, 2021, p. 42). Hence, it is probably not very difficult to imagine the threats for particularly qualitative and critical education research to operate under such conditions of dataveillance, more so, since it becomes more and more infused with shareveillance (Knox, 2019) – i.e., the norm of sharing one’s data, and to be rendered suspicious when denying to –, but equally with ethics. Again, while ethical accountability is undoubtedly key to all kinds of research, and particularly to education research, it is the infrastructuring logic of this ethics governance that is worrying here. More specifically, we can observe how ethics governance increasingly merges with governance by data management and is brought into the same logic of regulations, rubrics, and checklists, yet with the primary aim of securing data flow. This also means, however, that ethical accountability shifts from an ongoing process of critical (self-) reflection and professional peer control towards a second-order activity that can be administered, standardized, optimized, and at least partly outsourced (e.g., to ethical boards, or an anonymization technology, see also Robson & Malette, 2023). In doing so, it not only detaches ethics from the daily practices

of the researcher in the field, but equally, and more importantly, it diverts the gaze from – and makes it increasingly hard to challenge – the ethical problems that themselves lie within the seemingly technological-neutral infrastructures (see, for instance, the AI example above).

What to Do? Some Concluding Thoughts

The focus of this essay was on governance dynamics of research data infrastructuring, as they manifest in RDM discourses, policies, and programs today, and on their so far broadly underestimated risks for (particularly qualitative and critical) education research. This does not mean that there are no initiatives (e.g., data curation centers) that try to do things differently, for instance, through actively embracing the specific nature of qualitative education research. As an example, the German initiative QualiBi (www.qualibi.net) focuses a lot more on exchange facilitation between researchers, including forums for collective reflection on datasets and on a data platform adapted to the very specific needs of qualitative education research. At the same time, from a global perspective, such initiatives form, at least so far, an absolute exception in the education field, and they are challenged with the very same logic of infrastructuring (e.g., how to structure data platforms, which regulations for data submission to implement, where (not) to use existent standards, etc.). Still, they are undoubtedly an important part of (re)introducing more critical perspectives to the field, also because they start from adaptations *within* the infrastructural setting of governance. Such attempts to install alternatives within the broader systems of RDM or to produce what I would call “reflective cracks” in the infrastructure (for instance, through hampering smooth data flows) could indeed be thought further. One way could be to (re)establish elements of non-standardization (e.g., making use of free-text fields instead of predefined boxes to tick, see also Strickert, 2021, p. 65); another to deliberately use different designs in data platforms in order to illustrate how platform designs change what researchers see (differently) from education and education research. Further inspiration here can also be found in the field of so-called glitch studies that emphasize “[...] digital errors and failures in order to bring to the fore normative modes of engagement with digital technology” (Kemper & Kolkman, 2018, p. 2087). A third way could be to pay more attention to the aforementioned data intermediaries, in terms of the massive amount of labor happening at various parts of the infrastructure, but also in terms of their effects on educational ‘world-making’ and/or normalization. It is to be expected that such closer views will reveal a wide range of ambiguous practices, as an initial study by Strickert (2021) has already indicated.

At the same time, this essay would strongly call for not limiting critical perspectives to system-internal needs for adaptation. Rather, let us remind ourselves what education research, and particularly critical education research, is about, namely to expose how relations of power and inequality “are manifest and are challenged in the formal and informal education of children and adults” (Apple et al., 2009, p. 3), and to equally question what counts as legitimate knowledge/epistemological assumptions (Apple et al., 2009). From such a perspective, research data infrastructuring *must* also be tackled by more fundamental critique in the sense of a systematic deconstruction of the fine-grained mechanisms and dynamics of data infrastructuring.² Such critique needs solid investments in what Hartong and Sander (2021) called critical data(fiction) literacy, that is, researchers’ critical understanding of the politics and governmental dynamics of data management, as well as the wider infrastructural contexts they are participating in. This, obviously, requires different formats and a different tempo than training or checklists, as they are typically provided in RDM so far. It is through such strengthening of literacy that more fundamental resistance can fruitfully evolve, resistance which is not satisfied by attempts to make data

infrastructures fairer through adding more infrastructure of the same kind.

In the end, it may be useful to ask ourselves where problems such as ‘bad research’ – which shall become better controlled through ‘open science’ – are originating in the first place. Just as this question may, in some cases, indeed lead to scholars that take their profession not serious enough, in many other cases, it will lead to highly problematic structural research environments. Environments that are characterized by performance pressure on, and overload of, researchers, by the managerialization of research – including projectification, milestone plans, and the gradual elimination of unknown research results –, but foremost by a dramatic underfinancing of public research institutions. Interestingly, it is exactly the same conditions that seem to apply to the (crowd)work of many RDM intermediaries already today (Strickert, 2021), which makes the situation even more dramatic. It is to be expected that all these conditions will not improve but will further intensify if we let a system of fine-grained dataveillance become the default option.

Acknowledgement

This work was supported by the German Research Foundation [Project no. 423781123].

References

Apple, M. W., Au, W., & Gandin, L. A. (Eds.). (2009). *The Routledge international handbook of critical education*. Routledge.

BMBF (Bundesministerium für Bildung und Forschung) (2023). *Nationale Forschungsdateninfrastruktur*.
https://www.bmbf.de/bmbf/de/forschung/das-wissenschaftssystem/nationale-forschungsdateninfrastruktur/nationale-forschungsdateninfrastruktur_node.html

Bowers, A. J., & Choi, Y. (2023). Building school data equity, infrastructure, and capacity through FAIR data standards: Findable, Accessible, Interoperable, and Reusable. *Educational Researcher*, 52(7), 450–458.
<https://doi.org/10.3102/0013189X231181103>

Burgelman, J. C. (2021). Politics and open science: How the European Open Science Cloud became reality (the untold story). *Data Intelligence*, 3(1), 5–19.
https://doi.org/10.1162/dint_a_00069

Corrall, S. (2012). Roles and responsibilities: Libraries, librarians and data. In G. Pryor (Ed.), *Managing research data* (pp. 105–133). Facet Publishing.

Cox, A. M., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of librarianship and information science*, 46(4), 299–316.

Cox, A. M., & Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, 70(2), 142–157.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*.

Yale University Press.

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, 50(1), 101–113.

<https://doi.org/10.1111/bjet.12595>

Decuyper, M., Grimaldi, E., & Landri, P. (2021). Introduction: Critical studies of digital education platforms. *Critical Studies in Education*, 62(1), 1–16.

<https://doi.org/10.1080/17508487.2020.1866050>

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*, 2(2).

<https://doi.org/10.1177/205395171561718>

Edwards, R. (2015). Knowledge infrastructures and the inscrutability of openness in education. *Learning, Media and Technology*, 40(3), 251–264.

<https://doi.org/10.1080/17439884.2015.1006131>

Gerecht, M., & Kminek, H. (2018). Wie offen können und dürfen Forschungsdaten sein? *Erziehungswissenschaft*, 29(2), 9–10.

<https://doi.org/10.25656/01:16167>

Gewirtz, S., & Cribb, A. (2020). What works? Academic integrity and the research-policy relationship. *British Journal of Sociology of Education*, 41(6), 794–806.

<https://doi.org/10.1080/01425692.2020.1755226>

Giannoutakis, K. M., & Tzovaras, D. (2017). The European strategy in research infrastructures and Open Science Cloud. In *Data Analytics and Management in Data Intensive Domains: XVIII International Conference, DAMDID/RCDL 2016, Ershovo, Moscow, Russia, October 11-14, 2016, Revised Selected Papers* (pp. 207–221). Springer International.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. *Acm Sigmod Record*, 34(4), 34–41.

<https://doi.org/10.1145/1107499.1107503>

Hartong, S. (2016). Between assessments, digital technologies and Big Data: The growing influence of ‘hidden’ data mediators in education. *European Educational Research Journal*, 15(5), 523–536.

<https://doi.org/10.1177/147490411664896>

Hartong, S., & Förschler, A. (2019). Opening the black box of data-based school monitoring: Data infrastructures, flows and practices in state education agencies. *Big Data & Society*, 6(1).

<https://doi.org/10.1177/2053951719853311>

Hartong, S., & Sander, I. (2021). Critical Data(fication) Literacy in und durch Bildung. In A. Renz, B. Etsiwah, H. Burgueño & T. Ana, (Eds.), *Whitepaper Datenkompetenz* (pp. 19–20). Weizenbaum Institute for the Networked Society.

<https://doi.org/10.34669/wi/3>

Hartong, S., Machold, C., & Stosic, P. (2020). Zur (unterschätzten) Eigendynamik von Forschungsdateninfrastrukturen. Ein Kommentar zu den “Empfehlungen zur Archivierung,

Bereitstellung und Nachnutzung von Forschungsdaten im Kontext erziehungs- und bildungswissenschaftlicher sowie fachdidaktischer Forschung". *Erziehungswissenschaft*, 31(61), 51–59.

<https://doi.org/10.3224/ezw.v31i2.06>

Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51(7–8), 662–672.

<https://doi.org/10.1080/01930826.2011.601269>

Hirschauer, S. (2014). Sinn im Archiv? Zum Verhältnis von Nutzen, Kosten und Risiken der Datenarchivierung. *Soziologie*, 3(43), 300–312.

Huber, E. (2019). Affektive Dimensionen von Forschungsdaten, ihrer Nachnutzung und Verwaltung. *Working Papers des SFB 1171 "Affective Societies – Dynamiken des Zusammenlebens in bewegten Welten"*, 01/19, n.p.

<https://doi.org/10.17169/refubium-2481>

Imeri, S. (2017). Open Data? Zum Umgang mit Forschungsdaten in den ethnologischen Fächern. In J. Kratzke & V. Heuveline (Eds.), *E-Science-Tage 2017: Forschungsdaten managen* (pp. 167–178). heiBOOKS.

<https://doi.org/10.11588/heibooks.285.377>

Jornitz, S. (2008). Was bedeutet eigentlich "evidenzbasierte Bildungsforschung"? Über den Versuch, Wissenschaft für Praxis verfügbar zu machen am Beispiel der Review-Erstellung. *Die Deutsche Schule*, 100(2), 206–216.

<https://doi.org/10.25656/01:27247>

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 22(14), 2081–2096.

<https://doi.org/10.1080/1369118X.2018.1477967>

Kerssens, N., & Dijck, J. V. (2021). The platformization of primary education in The Netherlands. *Learning, Media and Technology*, 46(3), 250–263.

<https://doi.org/10.1080/17439884.2021.1876725>

Knox, J. (2019). What does the 'postdigital' mean for education? Three critical perspectives on the digital, with implications for educational research and practice. *Postdigital Science and Education*, 1(2), 357–370.

<https://doi.org/10.1007/s42438-019-00045-y>

Krammer, G., & Svecnik, E. (2020). Open Science als Beitrag zur Qualität in der Bildungsforschung. *Zeitschrift für Bildungsforschung*, 10(3), 263–278.

<https://doi.org/10.1007/s35834-020-00286-z>

Lewis, S., & Hartong, S. (2022). New shadow professionals and infrastructures around the datafied school: Topological thinking as an analytical device. *European Educational Research Journal*, 21(6), 946–960.

<https://doi.org/10.1177/1474904121100749>



- Macgilchrist, F., Allert, H., Cerratto Pargman, T., & Jarke, J. (2023). Designing postdigital futures: Which designs? Whose futures? *Postdigital Science and Education*, 1–12.
<https://doi.org/10.1007/s42438-022-00389-y>
- McLeod, J., Childs, S., & Lomas, E. (2013). Research data management. In P. Alison (Ed.), *Research methods in information* (pp. 71–86). Neal-Schuman.
- Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information services & use*, 37(1), 49–56.
<https://doi.org/10.3233/ISU-170824>
- Pels, P., Boog, I., Henrike Florusbosch, J., Kripe, Z., Minter, T., Postma, M., Sleeboom-Faulkner, M., Simpson, B., Dilger, H., Schönhuth, M., von Poser, A., Castillo, R. C. A., Lederman, R., & Richards-Rissetto, H. (2018). Data management in anthropology: the next phase in ethics governance? *Social Anthropology*, 26(3), 391–413.
<https://doi.org/10.1111/1469-8676.12526>
- Plantin, J. C. (2021). The data archive as factory: Alienation and resistance of data processors. *Big Data & Society*, 8(1).
<https://doi.org/10.1177/20539517211007510>
- Plantin, J. C., Lagoze, C., & Edwards, P. N. (2018). Re-integrating scholarly infrastructure: The ambiguous role of data sharing platforms. *Big Data & Society*, 5(1).
<https://doi.org/10.1177/2053951718756683>
- Poirier, L. (2021). Reading datasets: Strategies for interpreting the politics of data signification. *Big Data & Society*, 8(2).
<https://doi.org/10.1177/20539517211029322>
- Poole, A. H. (2015). How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science*, 15, 101–139.
<https://doi.org/10.1007/s10502-014-9236-y>
- Pryor, G. (Ed.). (2012). *Managing research data*. Facet Publishing.
- Rafiq, M., & Ameen, K. (2022). Research data management and sharing awareness, attitude, and behavior of academic researchers. *Information Development*, 38(3), 391–405.
<https://doi.org/10.1177/02666669211048491>
- Reh, S., Müller, L., Cramme, S., Reimers, B., & Caruso, M. (2021). Warum sich Forschende um Archive, Zugänge und die Nutzung bildungswissenschaftlicher Forschungsdaten kümmern sollten—historische und informationswissenschaftliche Perspektiven. *Erziehungswissenschaft*, 31(2), 5–6.
<https://doi.org/10.3224/ezw.v31i2.02>
- Renbarger, R., Adelson, J. L., Rosenberg, J. M., Stegenga, S. M., Lowrey, O., Buckley, P. R., & Zhang, Q. (2023). Champions of transparency in education: What journal reviewers can do to encourage Open Science practices. *Gifted Child Quarterly*, 67(4), 337–351.

<https://doi.org/10.1177/00169862231184575>

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, 67(12), 2964–2975.

<https://doi.org/10.1002/asi.23529>

Robson, K., & Malette, N. (2023). The ethics and bureaucratization of data management. In M. D. Young & S. Diem (Eds.), *Handbook of Critical Education Research*. (pp. 675–694). Routledge.

<https://doi.org/10.4324/9781003141464-40>

Sefton-Green, J., & Pangrazio, L. (2021). Platform pedagogies –Toward a research agenda. In *Algorithmic Rights and Protections for Children*.

<https://doi.org/10.1162/ba67f642.646d0673>

Strickert, M. (2021). *Spezifika und Herausforderungen qualitativer Daten im Forschungsdatenmanagement*.

<https://edoc.hu-berlin.de/handle/18452/24061>.

Uzwyszyn, R. (2022). Steps towards building library AI infrastructures: Research data repositories, scholarly research ecosystems and AI scaffolding.

<http://repository.ifla.org/handle/123456789/2062>

Veršić, I. I., & Ausserhofer, J. (2019). Social sciences, humanities and their interoperability with the European Open Science Cloud: What is SSHOC?. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 72(2), 383–391. <https://doi.org/10.31263/voebm.v72i2.3216>

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., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1), 1–9.

<https://doi.org/10.1038/sdata.2016.18>

Williamson, B. (2018). Silicon startup schools: Technocracy, algorithmic imaginaries and venture philanthropy in corporate education reform. *Critical studies in education*, 59(2), 218–236.

<https://doi.org/10.1080/17508487.2016.1186710>

Recommended Citation

Hartong, S. (2023). The global infrastructuring of research data and its risks for (particularly qualitative and critical) education scholarship. *On Education. Journal for Research and Debate*, 6(18).

https://doi.org/10.17899/on_ed.2023.18.7



Do you want to comment on this article? Please send your reply to editors@oneducation.net. Replies will be processed like invited contributions. This means they will be assessed according to standard criteria of quality, relevance, and civility. Please make sure to follow editorial policies and formatting [guidelines](#).

Sigrid Hartong

Sigrid Hartong is professor for sociology at the Helmut-Schmidt-University in Hamburg. Her research focus is on the transformation of governance in education and society, with a particular emphasis on the manifestations and effects of digitization and datafication.

1. https://www.forschungsdaten-bildung.de/files/Stellungnahme_zum_FDM_DGfE-GEBF-GFD.pdf
2. Even though this is not in the focus of this essay, such fundamental critique must equally address the environmental effects of large-scale data storage and processing.