




The *What* of Data: Defining Which Scientific Research Is Appropriate to Share

Bernadette M. Boscoe^(✉) 

University of Washington, Seattle, WA 98195, USA
boscoe@uw.edu

Abstract. Increasingly, scientists are releasing research data to the public for potential (re)use. Yet, the *what* of data—what gets shared (or kept private), by whom, and why—is difficult for data curators and stewards to determine. Scientific field-specific norms play an important role in decision-making processes to define what data are deemed acceptable to release. I explore the framework of *contextual integrity* (CI), which operationalizes appropriate flows of information that reflect context-dependent norms. CI is essentially a theoretical framework for privacy *in* data; however, in this work, appropriate data sharing *surrounds* the data. In this paper, CI methods are applied to a case study in astronomy and show how CI can guide an understanding of which data can be shared by tracing how people move information within contexts. The aim is to provide both researchers and repository maintainers an approach to make data available in an appropriate way that does not violate rapidly evolving sharing norms.

Keywords: Data · Sharing · Infrastructures

1 Introduction

In compute-intensive science, the availability of data has skyrocketed in recent years, paired with an interest in making these data available for potential (re)use. Much scholarly work has sought to understand the complex landscape of incentives, restrictions, and policies within open science that both support and limit data sharing [16]. Shifts in data access and control are occurring in scientific research fields as data are increasingly seen as being valuable and, therefore, worthy of greater attention. In this paper, I limit my inquiry to a narrow aspect: Assuming researchers are willing to make their research data available, what are “the data” that can be appropriately shared so as to best reflect community norms? In scientific research processes, “the data” are impossible to define: data enjoy numerous evidentiary and statistical iterations throughout the research

Supported by the Alfred P. Sloan Foundation, grant numbers 201811217 and 201514001. Special thanks to Nic Weber.

process and are contingent upon their eventual purpose [5]. Previous work has explained the circulation of data as a commodity governed by both property rights and contextual social norms [17]. In what follows, I focus on the latter, norms-based governance, through the framework of contextual integrity (CI), a descriptive and diagnostic framework that makes it possible to judge appropriate flows of information given context dependent norms. Many applications of CI have focused on normative violations that involve data content of a sensitive nature (e.g. personally identifiable information (PII)), however the strength of CI is that it provides a comprehensive means of analyzing not just the content of data, but the broader sociotechnical environment in which data are shared. Here, I apply CI to a case study in astronomy. This work aims to surface relationships between norms and data to best inform the design of relevant infrastructures for making research data available for potential reuse within spaces where data sharing is contested.

2 Data to Share

More than ever, digital infrastructures such as data repositories offer researchers opportunities to make their data publicly available. Researchers' motivations and disincentives to share and reuse research data have been explored by a number of scholars in information science fields [3, 13, 14]. Reproducibility rises to the fore as a motivation to share for purposes of scientific rigor, but the laborious nature of its implementation can be a challenge for scholars [7, 15]. At the same time, individual researchers can receive greater visibility for their work by relying on open-science tools and linkages, perhaps leading to more citations and thus justifying efforts to make data more available to the public [6]. Motivations aside, should researchers decide to make their data available, the question of *what* to make available involves choosing and organizing data, metadata, and code, drawing from—in essence—a flowing stream of information that can be placed into a data repository. Sharing proves more challenging for small research groups who lack enterprise-scale support structures and staff to assist in data sharing practices.

3 Contextual Integrity *Around* Data

CI is a framework developed by Nissenbaum et al. [10, 12] that defines *privacy* as appropriate information flows based on norms specific to contexts. CI was developed for use with data containing information about individuals and has been adopted by researchers studying the Internet of Things and social media, to name a few examples [2]. Five parameters characterize information flow: subject, sender, recipient, information type, and transmission principles. Senders and recipients are actors that can be individuals or groups of people. Information type is a description of the form information takes, for example, an email. Transmission principles are defined as constraints imposed upon the information flows. Privacy is breached when an information flow fails to map onto expected

values. Importantly, to ascertain privacy norms when evaluating a scenario, all five parameters must be specified. Some transmissions can be a violation in one sphere yet not in others. Additionally, transmission principles that change over time can result in violations from previously acceptable flows.

The main purpose of CI is to examine information flows that contain data about individual privacy; however, researchers can also violate data sharing norms by behaving in ways deemed unacceptable to others in their field (or other fields). In the research world, information is a highly coveted commodity—and the published results are often rewarded for being novel and first. Nissenbaum’s [11] new work employs the metaphor of a *data food chain*, a hierarchical construct where “data of a higher order are a function of data of a lower order” (p. 236) thus stratifying data into layers that can be mapped to effects on privacy. Nissenbaum [11] notes that *data primitives*, which are event imprints such as electrical signals or activated pixels or GPS coordinates, are challenging to map onto norms; it is the higher levels of semantic data that can be more easily evaluated. For example, a mouse “click” is a meaningless trace until it is put into the context of higher-level data-processing layers. Thus, CI cannot be applied at the click level. In acknowledging the complexities of the data layers and related actors and norms, I turn this problem on its head and consider researchers, with their layers of data and multiple normative spheres through which to navigate, as a way of evaluating that information which is acceptable to share.

4 Data Sharing in Astronomy

Astronomy (including astrophysics, cosmology, and related fields) is a digital data-rich field with a multitude of data sources from observational and theoretical domains, often one serving to verify one another. Observational instruments collecting data take many forms, from ground and space telescopes to weather balloons and radio dish arrays. Compared with other science fields, astronomy has a well-established history of sharing data with the public [8]. However, a more granular inspection of data sharing practices reveals differences between countries, sub-fields, locations, projects, and instruments, as well as between PIs and among their teams. Microsoft’s database guru, the late Jim Gray, famously said that his interest in astronomy data stemmed from the fact that it had no commercial value [9]. The ones and zeros that make up digital explorations into the night sky may not possess a market value, but the scores of scientists, data wranglers, archivists, instrument operators, and others working within the realm certainly are involved in the ethical entanglements that are part and parcel in doing science.

4.1 A Brief History of Norms of Astronomy Data Sharing

In the 1970s, acquiring astronomy data was a physical act and a heavy one at that: Astronomers had to lug reels of tapes from mountaintop observatories and drag heavy suitcases of data into cars and onto planes to return home

and begin analyzing their data. By the 1990s, the digitization of astronomy was nearly complete, the Internet might not have been capable of transmitting astronomically large amounts of data, but media in the form of hard drives or discs made it easier for astronomers to share. In the United States, a plethora of telescopes were in the process of being built on Earth or launched into space, such as Hubble (1990), Chandra (1999), Spitzer (2003), and Keck (1992). At this time, U.S. space missions were generally funded by NASA and ground-based telescopes tended to be funded by private organizations. As a result, data sharing norms differed widely between the privately and publicly funded operations and between national and international entities [18]. Hubble Space Telescope is a canonical example of NASA data sharing practices. From the initial planning stages, Hubble data were to be put in an archive made available to the public. Levels of data reductions were categorized, with Level 0 data being “raw” data from the telescope, and subsequent numbers indicating that more processing had been done to the data. To collect Hubble data, astronomers submit proposals to observe various phenomena in the night sky, which are then processed by Hubble staff and given to the observer team. The observer is typically given a grace period of 12–18 months, offering them the opportunity to develop and publish results; after this time, the data are made available to the public.

It is important to stress that data released in large, mission-based public archives differs from research data created in the act of doing science as a result of analysis to be used for publications. The latter form of data may be shared by being placed on team or individual repositories, on university servers, or within platforms such as Zenodo or GitHub. Often, the research datasets that are made available with publications are voluntarily determined by the PI of the project, making them especially nebulous and unpredictable across cases.

4.2 Methods

To show examples of CI frameworks demonstrating data sharing practices, I draw from findings from a three-year qualitative case study of astronomy [4]. I conducted ethnography and participant observation at six locations, including observatories and universities with astronomy research groups, interviewing 40 astronomers and other related staff, such as programmers, and data repository stewards. The findings also stem from a corpus of interviews of astronomers done by information scientists within the Center for Knowledge Infrastructures at UCLA, led by Dr. Christine Borgman. I also spent two years embedded in an astronomy research group studying their data and code practices. A detailed understanding of data sharing practices grew from this in-depth and up-close scrutiny. Three vignettes follow from discussions with astronomers in the case study.

4.3 Scenario 1: “Horse Trading” as a Sharing Mechanism Between Teams

{*Subject*: Observational data. *Sender*: Japanese research team. *Receiver*: U.S. research team. *Information type*: Analyzed, reduced data from the Japanese telescope. *Transmission principle*: Data to be shared in between the two teams only; not explicitly stated but intended for the single use of producing one collaborative research paper}.

In 2018, members of a Japanese astronomy research group were working on a project about stars in the Milky Way Galaxy similar to that of a U.S.-based research team. The Japanese team sent a long formal email letter to the U.S. team requesting to share data and code with each other. In return for the U.S. team’s data, the Japanese team would give the U.S. team their data plus authorship on a related publication. After some consideration as to whether it was beneficial to do so, the U.S. team agreed to the exchange. Recall, transmission principles are constraints imposed upon the information flow; so in this case the data shared was to remain private between the two groups. Therefore, if the U.S.-team were to share this data with a third party or place it into a public repository without the Japanese group’s permission, this would be considered a norm violation.

This type of *horse trading* (a term used by astronomers) predates newer scenarios to make research data available with an associated publication. I recommend that a data steward, such as an archivist or repository manager, be made aware of the provenance of various datasets, as well as any associated understandings and agreements, especially because assumptions might not have been explicitly stated. Some research groups in astronomy choose to share all data and code associated with their research, but others are forbidden from doing so by their PIs. To borrow from mathematical set theory, these datasets might be thought of as *clopen*—both open and closed and thus requiring more care in terms of curation. Repository designers’ implementations might evolve as a result of surfacing norms via CI approaches. So, too, can CI be used to formalize design requirements in a way that can be translated into technical solutions [1].

In follow-up interviews, the U.S. team said they did end up sharing with the Japanese team, and the results were published in a journal that required the data to be made available to the public. However, it was unclear whether it would be acceptable to use the Japanese data for further projects, even though the data are public. Questions of ownership and appropriate use remain. One collaborator with the U.S. team said, when asked what an infrastructure could look like to improve sharing, “If data is shared for a one-time use, you could put a key on it so that it expires”. He continued to explain his own relationship with the U.S. team data and said, “I must respect the data. I can access it but cannot use it unless the [U.S.] PI says yes”. Another team member of the U.S. team opined that “We need a real data release policy beyond the PI informally deciding what gets shared. We want things to be clearer”.

4.4 Scenario 2: Changing Transmission Principles

{*Subject*: Exoplanets. *Sender*: PI of research group. *Receiver*: Astronomy journal. *Information type*: DOI and links to repository. *Transmission principle*: Must include DOI and links in a particular way}.

CI can reveal changing transmission principles by comparing parameters over time and by examining norms. The following example demonstrates this. In September 2019, an astronomer in a small research group discussed reviews of her article recently accepted by an astronomy journal. She explained, “This is the first time I have ever seen explicit instructions for how to share the data associated with this paper. I share my data in GitHub, and the reviewer told me to get a DOI from Zenodo and link to my GitHub repository and use the DOI in my citations”.

This is an interesting constraint imposed upon the researchers: As a condition to publish, they must share data in a way specified by the journal. Whether this is a norm violation is contingent upon the person or group evaluating this case; for all intents and purposes, the introduction of new contingencies placed on sharing practices is of importance to the repository designer and maintainer; knowing that researchers must share in a specific way allows for codification of a set of rules with which the repository can align. This case study also demonstrates potentially clashing norms of research groups, publishers, repositories, and even prior field norms that have dictated previous sharing practices.

4.5 Scenario 3: Violating the “Gentleperson’s Agreement”

{*Subject*: Data in an archive. *Sender*: A rival research team. *Recipient*: A journal. *Information type*: A paper and associated data for publication. *Transmission principle*: Paper submitted to this journal and not elsewhere}.

I asked astronomers to give an example of a norm violation that might be construed as egregious across the board. An astronomer explained the following case: “Let’s say a team took a series of observations and were working on a paper. The embargo period passes, and the raw data is released into the public archive but the team isn’t finished with the paper yet. Another group uses the data from the archive to publish the same result first, scooping the other team”. Other astronomers agreed that was “not cool” and that it is a problematic practice. I asked what would happen in this case, and astronomers replied that nothing would really be done, but that the offending team might get a bad reputation for doing this. “It is just something that should not be done” opined another astronomer, and added, “We call them archive vultures”.

5 Conclusion

As fields such as astronomy have amassed increasing amounts of data and are fostering larger than ever collaborations among research groups, informal forms of sharing break down. Research repositories are but one way to make rules

of reusing data more explicit, especially within public frameworks. A better understanding of what data gets shared can result in improved infrastructures to promote appropriate reuse. I have shown that by using the framework of CI, potential norm violations in the transmitting of information can occur in many situations of data sharing. Instead of looking at a dataset as an entity that can or cannot be shared, I instead evaluate its transmission within the various contexts in which it might be shared. A deeper understanding of these nuances can better inform the design and maintenance of repositories tasked with sharing data with different rules attached. As data sharing in science increases, creating more sophisticated ways to share data on different levels will help address the problem of what data can be shared and with whom. In particular, understanding who makes these decisions within a science field has a pronounced effect on what data get shared. As field norms shift over time, repository designers can enable functionality to determine what data to share.

References

1. Barth, A., Datta, A., Mitchell, J.C., Nissenbaum, H.: Privacy and contextual integrity: framework and applications. In: 2006 IEEE Symposium on Security and Privacy (S P 2006). 15pp.–198, May 2006. <https://doi.org/10.1109/SP.2006.322>
2. Benthall, S., Gürses, S., Nissenbaum, H.: Contextual integrity through the lens of computer science. *Found. Trends Priv. Secur.* **2**(1), 1–69 (2017). <https://doi.org/10.1561/33000000016>
3. Borgman, C.L.: *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, Cambridge (2015). <http://mitpress.mit.edu/big-data>
4. Boscoe, B.M.: *From blurry space to a sharper sky: keeping twenty-three years of astronomical data alive*. Ph.D. thesis, UCLA (2019). <https://escholarship.org/uc/item/2jv941sb5>
5. Collins, H.M.: The meaning of data: open and closed evidential cultures in the search for gravitational waves. *Am. J. Sociol.* **104**(2), 293–338 (1998). <https://doi.org/10.1086/2100406>
6. Curty, R.G., Crowston, K., Specht, A., Grant, B.W., Dalton, E.D.: Attitudes and norms affecting scientists' data reuse. *PLoS One* **12**(12), e0189288 (2017). <https://doi.org/10.1371/journal.pone.0189288>
7. Darch, P.T., Borgman, C.L.: ShipSpace to database: motivations to manage research data for the deep seafloor biosphere. In: Proceedings of the 77th Annual Meeting of the Association for Information Science and Technology, Seattle, WA, November 2014. <http://www.asis.org/asist2014/proceedings/submissions/papers/156paper.pdf,000008>
8. Genova, F.: Data as a research infrastructure CDS, the Virtual Observatory, astronomy, and beyond. In: EPJ Web of Conferences, vol. 186, p. 01001 (2018). <https://doi.org/10.1051/epjconf/201818601001>
9. Hey, T.: The big idea: the next scientific revolution. *Harv. Bus. Rev.* (2010). <https://hbr.org/2010/11/the-big-idea-the-next-scientific-revolution10>
10. Nissenbaum, H.: Privacy as contextual integrity. *Wash. Law Rev.* **79**, 39 (2004)
11. Nissenbaum, H.: Contextual integrity up and down the data food chain. *Theor. Inq. Law* **20**(1), 221–256 (2019). <https://www7.tau.ac.il/ojs/index.php/til/article/view/161412>

12. Nissenbaum, H.F.: *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, Stanford (2010)
13. Palmer, C.L., Weber, N.M., Cragin, M.H.: The analytic potential of scientific data: understanding re-use value. *Proc. Am. Soc. Inform. Sci. Technol.* **48**(1), 1–10 (2011). <http://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801174/full14>
14. Pasquetto, I.V., Randles, B.M., Borgman, C.L.: On the reuse of scientific data. *Data Sci. J.* **16** (2017). <https://doi.org/10.5334/dsj-2017-008>. <http://datascience.codata.org/articles/10.5334/dsj-2017-008/0000015>
15. Stodden, V., et al.: Enhancing reproducibility for computational methods. *Science* **354**(6317), 1240–1241 (2016). <https://doi.org/10.1126/science.aah6168>
16. Tenopir, C., et al.: Data sharing by scientists: practices and perceptions. *PLoS One* **6**(6), e21101 (2011). <https://doi.org/10.1371/journal.pone.0021101>
17. Vertesi, J., Dourish, P.: The value of data: considering the context of production in data economies. In: *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW 2011*, pp. 533–542. ACM, New York (2011). <https://doi.org/10.1145/1958824.1958906>
18. Williams, T.: The philanthropy of stargazing: we're in a new golden age of mega telescope projects (2014). <http://www.insidephilanthropy.com/home/2014/12/18/the-philanthropy-of-stargazing-were-in-a-new-golden-age-of-m.html>