OPEN WITH PURPOSE
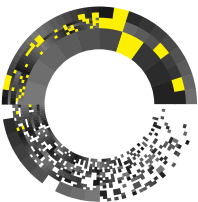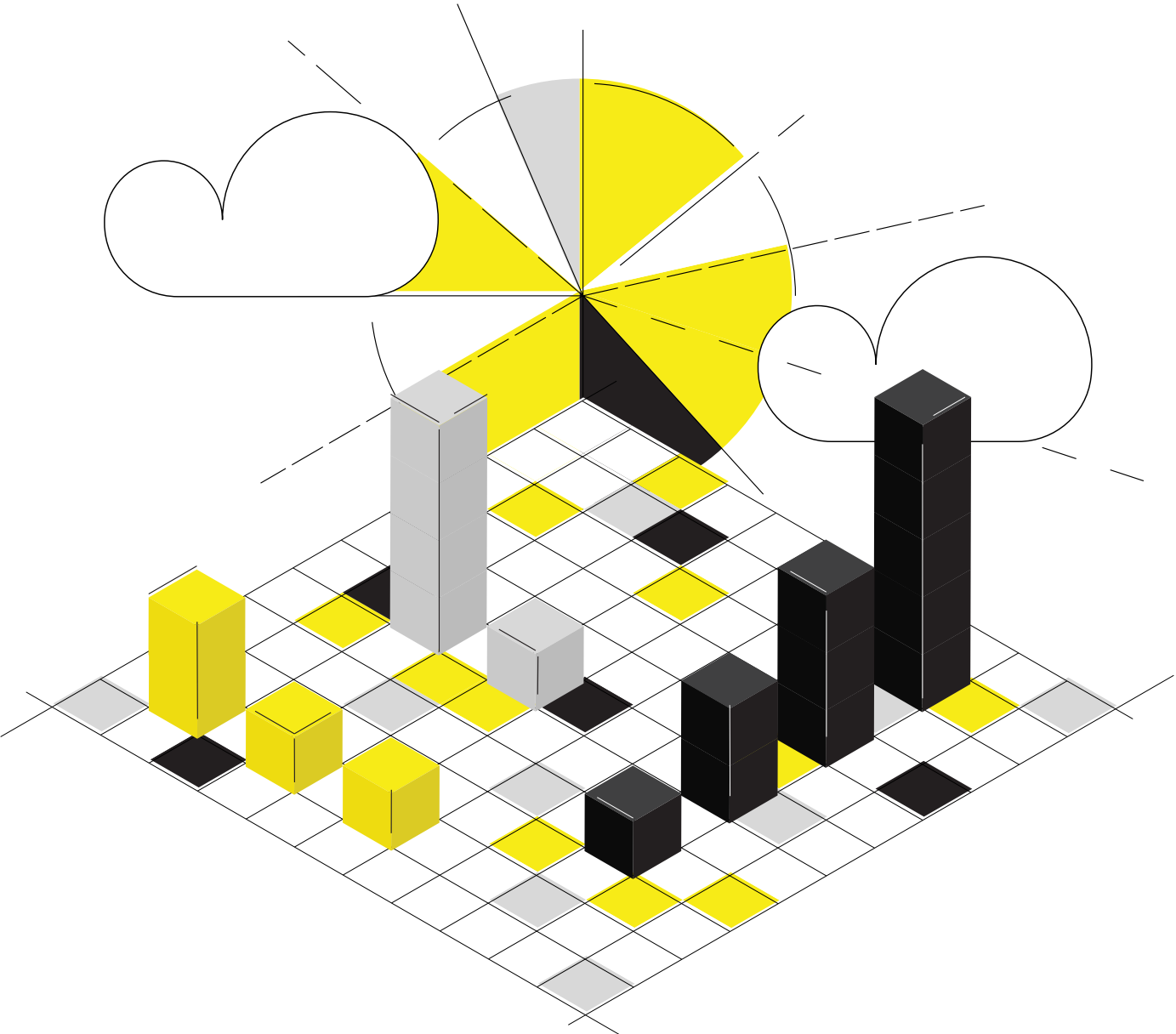
Ethical Issues and Stewardship Mechanisms for Open Climate Data

OPEN
_FUTURE

JUNE 2024

Aditya Singh

INTRODUCTION

There is increasing interest in data-driven interventions to the climate crisis, and in particular, in Open Data. A recent report by the World Resources Institute concluded that Open Data could have the following benefits in the context of climate action: improved data coordination and quality, better-informed decision-making; greater coordination and novel partnerships, democratizing modeling approaches; enhanced monitoring of policies and programs.[1] While there exists no universally accepted definition of Open Data, it is typically understood as data "that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere."[2]

This report reviews both academic and "grey" literature on Open Data, with a view toward identifying key ethical challenges and issues that collecting and processing such data in the context of climate action might raise.

This review draws from broader critiques of data infrastructures to highlight the particular challenges of creating open-data collections and platforms. Finally, the report explores mechanisms and practices that open-data projects may put in place in order to mitigate some of these risks. This review does not focus on the technical and operational challenges that the creation and maintenance of open-data collections and platforms may present.[3]

---

[1] Grinspan, Delfina, and Jesse Worker. "Implementing Open Data Strategies for Climate Action: Suggestions And Lessons Learned for Government and Civil Society Stakeholders." World Resources Institute, 2021. doi:10.46830/wriwp.19.00093.

[2] The Open Data Charter, https://opendatacharter.net/principles/; See Verhulst, Stefaan, and Andrew Young. Global Impact of Open Data. [Place of publication not identified] O'Reilly Media, 2016 for a list of alternative definitions.

[3] For instance, challenges related to data standards, data quality, or compliance with regulatory frameworks. See Zuiderwijk, Anneke, Marijn Janssen, Sunil Choenni, Ronald Meijer, and Roexsana Sheikh Alibaks. "Socio-technical Impediments of Open Data." Electronic Journal of E-Government 10, no. 2 (December 1, 2012): pp156-172-pp156-172. https://academic-publishing.org/index.php/ejeg/article/view/571 for a discussion on various social and technical impediments to Open Data.

ISSUES AND RISKS

De-anonymization and Identifiability

Data sets shared as Open Data are usually anonymized, aggregated, or scrubbed of any identifiable characteristics. However, Open Data may still raise risks related to privacy and identification. Once released to the public, data cannot be withdrawn, and data may still become identifiable in combination with additional data sets. In addition, data analytics techniques are undergoing constant improvements, and the computing power available for analysis is also receiving constant investment. Thus, if data represents human behavior in any way, it can pose ongoing risks from the perspective of data protection, privacy, and identifiability. This is especially the case if the objective of an open-data platform is to collect, index, and aggregate more and more data sets, and also provide analytic capabilities.[4]

A now famous case from 2006 is Netflix's release of records revealing hundreds of thousands of user ratings from 1999 to 2005 and offering a prize to teams that could improve their movie recommendation algorithm. Within weeks, researchers were able to re-identify a subset of people by cross-referencing the Netflix data with IMDB.com ratings. The researchers were able to re-identify individuals using six ratings of obscure movies if they were found in both data sets. If the researchers included the approximate time the rating was made, they were able to re-identify users 99% of the time. Other studies have found that a combination of data points such as gender, birth date, and postal codes can often be sufficient to re-identify individuals.[5]

In a more recent study related to environmental health data, researchers evaluated the privacy risks related to sharing data sets from twelve prominent studies.[6] They found that data sets from all twelve studies included at least two of five selected data types[7] that could be found in publicly or commercially available data sets, raising re-identification risks. For one study, participants' region of residence could also be inferred with 80%–98% accuracy using environmental measurements found in the data. The researchers noted that various group markers, such as race or gender, could also likely be inferred from the studies and that similar outcomes (for individual and group identifiability) have previously been found for other types of data, such as bicycle use, electricity meter readings, and hospital visits.

---

[4] Weber, Nic, and Brandon Locke. "Ethics of Open Data," 2022. doi:10.48550/ARXIV.2205.10402.

[5] resource-centre. "Resource-Centre." Accessed June 5, 2024. https://www.d4d.net/state-of-open-data/chapters/sectors/geospatial/.

[6] Boronow, Katherine E., Laura J. Perovich, Latanya Sweeney, Ji Su Yoo, Ruthann A. Rudel, Phil Brown, and Julia Green Brody. "Privacy Risks of Sharing Data from Environmental Health Studies." Environmental Health Perspectives 128, no. 1 (January 2020): 017008. doi:10.1289/EHP4817.

[7] These included data related to geographic location, medical data, occupation, housing characteristics, and genetic data.

These concerns are particularly salient when data is collected in contexts where there may not be data protection frameworks and protections in place. For data sets that are made open in such contexts, it may then be crucial to have internal safeguards and governance strategies in place.

Risks from Aggregated and Anonymized Data

Identification and disclosure of personal information relating to individuals is only one form of ethical risk that may arise from the sharing and opening up of data. In the era of big data, analytics are developed to operate at as broad a scale as possible, where the individual may only be incidental to the analysis. Data-related harms may arise not just from accessing personally identifiable information, but where inferences are drawn (for instance, by authorities or corporations) at the level of groups.[8]

Big data is based on identifying patterns, and profiling large and undefined groups, and not individuals. People are no longer targeted as individuals but as members of specific groups - as owners of a particular car, consumers of a particular product or service, fans of a specific genre of music, cat owners or dog owners, or living in a certain postal code.[9] The results of such grouping and analysis are also applied on a large, group-level scale through policies or other interventions, for instance, in the context of urban design, health, or disaster response.[10] Thus, harm may arise at the collective or group level without any need for specific individuals to be identified.

Many forms of social sorting and discrimination may not require personally identifiable information. Even when individuals are not identifiable, they may still be reachable i.e., subject to consequential inferences and predictions taken on that basis.[11] For instance, the average price of housing in a certain postal code would not be considered personal information. However, the average price of houses in a particular neighborhood, when matched with an individual's address, could provide an estimate of the value of a particular house. This can be consequential information in many contexts. Climate and agricultural data may often be geotagged or easily combinable with other forms of geographic and geospatial data, raising similar concerns.

Climate-related data sets can provide insight into land use, soil health, air quality, water levels, etc. This information can amplify existing asymmetries, for instance, if incorporated into

---

8 Barocas, Solon, and Helen Nissenbaum. "Big Data's End Run around Anonymity and Consent." In Privacy, Big Data, and the Public Good, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 1st ed., 44–75. Cambridge University Press, 2014. doi:10.1017/CBO9781107590205.004.

9 Floridi, Luciano. "Open Data, Data Protection, and Group Privacy." Philosophy & Technology 27, no. 1 (March 2014): 1–3. doi:10.1007/s13347-014-0157-8.

10 Taylor, Linnet, Luciano Floridi, and Bart Van Der Sloot. "Introduction: A New Perspective on Privacy." In Group Privacy, edited by Linnet Taylor, Luciano Floridi, and Bart Van Der Sloot, 1–12. Cham: Springer International Publishing, 2017. doi:10.1007/978-3-319-46608-8_1.

11 Barocas, Nissenbaum, "Big Data's End Run around Anonymity and Consent.".

decision-making models in contexts such as housing and land investments or insurance premiums. Borgesius et al[12] propose a related example, such as the release of open-data crime statistics by a local council. Vendors of GPS systems may incorporate this data into their own maps, designating certain areas as high-crime neighborhoods and routing drivers around them whenever possible. This practice risks amplifying asymmetries and deprivation in those areas if businesses may not receive as much footfall or the area may receive lesser investment overall. Insurance premiums might rise, and real estate prices and business profits may drop. Insights related to environmental conditions, air quality, soil and water health can be similarly crucial in resource allocation decisions. These risks may also be amplified if open-data platforms collate data from multiple sources and also house capabilities that enable further aggregation and analysis.

Bias Embedded in Data Sets

There is no such thing as "raw" data. (Open) data is often the by-product, and not the end-product, of certain activities, especially for data sets on platforms that would be generated from development and climate-related projects. Many data sets are also constructed in the process of being shared and made openly available.[13] Additionally, what information is recorded as data is not a natural phenomenon but a design choice that reflects the purposes, resources, and values of the data collector.[14]

The construction of data is a social process and may, therefore, embed within data sets existing social asymmetries, even as an unintended consequence. For data sets that aren't strictly about environmental conditions (e.g., air quality data) but represent people and human behavior, opening up data sets that comprise, these asymmetries may further propagate or amplify them. "Datized moments" i.e., instances where an interaction or event is encoded as data, typically occur in the interaction of an individual with a data-collecting entity such as a state, business, or even a non-profit entity. However, people and groups may differ in their propensity to interact with such organizations.

As a consequence, data may overrepresent some groups and underrepresent others. This difference often relates to existing structures and patterns of social privilege. For instance, migrants, indigenous groups, undocumented workers, or refugees may have lower trust in government and state institutions, eschewing interactions with data-collecting bureaucracies.

---

12 Zuiderveen Borgesius, Frederik, Mireille van Eechoud, and Jonathan Gray. "Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework." SSRN Scholarly Paper. Rochester, NY, November 24, 2015. https://papers.ssrn.com/abstract=2695005.

13 Davies, Tim, and Mark Frank. "'There's No Such Thing as Raw Data': Exploring the Socio-Technical Life of a Government Dataset'. In Proceedings of the 5th Annual ACM Web Science Conference, 75–78. WebSci '13. New York, NY, USA: Association for Computing Machinery, 2013. https://doi.org/10.1145/2464464.2464472.

14 Johnson, Jeffrey Alan. "From Open Data to Information Justice." Ethics and Information Technology 16, no. 4 (December 2014): 263–74. doi:10.1007/s10676-014-9351-8.

Johnson[15] references an example from the US Census, where minority households were often undercounted: "Households are not missed in the census because they are black or Hispanic. They are missed where the Census Bureau's address file has errors; where the household is made up of unrelated persons; where household members are seldom at home; where there is a low sense of civic responsibility and perhaps an active distrust of the government; where occupants have lived but a short time and will move again; where English is not spoken; where community ties are not strong."

This may also be the case for data that doesn't directly represent populations and communities. For instance, data sets about air quality in a particular region bear the imprint of unequal access to, and distribution of, equipment measuring such data. These may map onto existing socio-economic asymmetries, introducing bias in any interventions based on this data.

Thus, data sets that are made openly available may contain biases and patterns of social privilege from the contexts in which they were constructed. These biases may be exacerbated in contexts of greater income inequalities or digital divides. This suggests the requirement for having mechanisms in place to assess and audit these data sets for these asymmetries. Notably, Johnson[16] suggests that the presence of these biases and injustices in data sets is not in itself an argument against opening up data sets. The data sets being opened may be the only reason these biases may come to light in the first place.

Asymmetric Benefit and Impact

Open data may not invariably lead to positive outcomes, as they are contingent on the capacities and resources of actors to meaningfully benefit from the data.[17] Gurstein[18] describes seven components that enable potential data users to benefit: "1. Sufficient internet access that data can be accessed by all users. 2. Computers and software that can read and analyze the data. 3. Computer skills sufficient to use them to read and analyze data. 4. Content and formatting that allows use at a variety of levels of computer skill and linguistic ability. 5. Interpretation and sense-making skills, including both data analysis knowledge and local knowledge that adds value and relevance. 6. Advocacy in order to translate knowledge into concrete benefits. 7. Governance that establishes a regime for the other characteristics." This follows a general trend where the primary beneficiaries of digital tools and interventions skew toward middle-class and

---

[15] Johnson, Jeffrey Alan. "From Open Data to Information Justice."

[16] Johnson, Jeffrey Alan. "From Open Data to Information Justice."

[17] Beer, Jeremy De. "Ownership of Open Data: Governance Options for Agriculture and Nutrition," 2017. doi:10.7490/F1000RESEARCH.1114298.1.

[18] Gurstein, Michael B. "Open Data: Empowering the Empowered or Effective Data Use for Everyone?" First Monday, January 23, 2011. doi:10.5210/fm.v16i2.3316.

well-educated groups.[19] Thus, it may often be the case that only highly resourced stakeholders benefit from open data. An example is the COINS database released by the United Kingdom's Treasury.[20] After its release, the Treasury itself conceded that given its volume and technical complexity, the data required considerable expertise to analyze it. Even established and well-resourced news institutions such as the Guardian and BBC were unable to immediately make use of it, and interest in the usage of the data eventually dropped.[21]

In recognition of the importance of capacity and resources in translating Open Data into outcomes, the Open Data Charter in 2018 also announced a shift in focus from an approach of "publish, and they will come" towards "publishing with a purpose." There was a recognition that simply publishing as much data as possible, as fast as possible, may not result in large-scale impact.

There may also be asymmetries in benefits across geographies. Collaborations and projects between the institutions in the Majority World and the Minority World are often funded by the latter. There are often concerns that the flow of data and research is from the Majority World to the Minority World and rarely vice versa. For instance, scientists from African nations have often expressed concern that Open Data reopens the gates for "parachute research" i.e., researchers from the Minority World benefiting from data collected by others for their own career advancement without engaging with or benefitting the institutions or communities that collected the data, or to whom that data relates.[22] While the objective of Open Data is to enable a greater number of stakeholders to address societal problems, opening up data also results in researchers using and interpreting data from communities to which they have no connection.

Not all projects might necessitate local expertise, and research findings may not always have direct implications for community well-being. At the same time, researchers using Open Data should be encouraged to conduct studies that are ethically sound and relevant to the local communities and prioritize finding ways of returning benefits and results to the community.

Amplifying Existing Asymmetries

There is also growing recognition among open-data advocates that opening up data responsibly requires consideration for its impact on vulnerable communities. Opening up informational

---

[19] Davies, Tim, and Fernando Perini. "Researching the Emerging Impacts of Open Data: Revisiting the ODDC Conceptual Framework." The Journal of Community Informatics 12, no. 2 (June 21, 2016). doi:10.15353/joci.v12i2.3246.

[20] The Combined Online Information System known as COINS. COINS is the UK Treasury's main database for managing budgetary and expenditure data, and covers organizations across the public sector.

[21] McClean, Tom. "Not with a Bang but a Whimper: The Politics of Accountability and Open Data in the UK." SSRN Scholarly Paper. Rochester, NY, 2011. https://papers.ssrn.com/abstract=1899790.

[22] Serwadda, David, Paul Ndebele, M. Kate Grabowski, Francis Bajunirwe, and Rhoda K. Wanyenze. "Open Data Sharing and the Global South—Who Benefits?" Science 359, no. 6376 (February 9, 2018): 642–43. doi:10.1126/science.aap8395.

resources can mean exposing them to existing power structures, and it may be necessary to re-examine the assumption that opening up resources predominantly results in emancipatory and empowering consequences.[23] Data can represent knowledge that, if made open, can enable some stakeholders to benefit at the expense of others. Thus, Open Data, under conditions of unequal capabilities, can risk simply "empowering the already empowered".[24]

A commonly referenced example on this point, is the case of a program digitizing and opening up data about land records in Bengaluru[25] (a city in the South of India[26]). The digitization process ended up excluding the land claims of Dalit[27] groups, who typically did have evidence for their claims but not according to the formats and specifications of the program that was introduced. Eventually, real estate developers became the main beneficiaries of the program, as they were better positioned to gain access to and use digitized records. They had greater computational capabilities and advocacy resources in relation to the political and legal practices governing land tenure. This led to mass evictions of residents from areas that were profitable for developers, where these communities had previously had the ability to contest the developers' claims.

Land registers, surveys, and maps, now augmented with satellite imagery, GPS, and GIS data, can become forms of knowledge that are difficult for local people to access. This enables the diffusion of corporate power over local land and people by dissociating information regarding land from their social contexts and serving to marginalize local people.[28] Access to data may make local land and resources more legible and more investible to powerful interests, often at the expense of local populations. Thus, there is a risk that Open Data may only make structurally inequitable systems work more efficiently or reinforce existing social stratifications that underlie digital access.

In the context of agriculture in particular, there is increasing recognition that farm- and land-related data can be very valuable to businesses such as seed and chemical firms, agronomists, insurance providers, and technology companies. This is illustrated by ongoing investments that technology and seed providers have made in the sector. For instance, in 2012-13, Monsanto acquired Precision Planting Inc., an agricultural software and hardware company, for $210 million, and the Climate Corporation, a climate monitoring company, for $930 million[29]. In 2017,

---

[23] Keller, Paul, and Alek Tarkowski. "The Paradox of Open." Open Future, March 5, 2021. https://openfuture.pubpub.org/pub/paradox-of-open/.

[24] Gurstein, "Open Data: Empowering the Empowered…".

[25] Previously named Bangalore.

[26] See Gurstein, "Open Data: Empowering the Empowered…" for a summary of the case of land records digitization.

[27] Historically marginalized caste groups in the Indian subcontinent.

[28] Fraser, Alistair. "Land Grab/Data Grab: Precision Agriculture and Its New Horizons." The Journal of Peasant Studies 46, no. 5 (July 29, 2019): 893–912. doi:10.1080/03066150.2017.1415887.

[29] Kesmodel, David. "Monsanto to Buy Climate Corp. for $930 Million". Wall Street Journal, 2 October 2013, sec. Business. http://online.wsj.com/article/SB10001424052702304176904579111042198455468.html, access date: 04.02.2024.

John Deere paid $305 million to acquire Blue River Technology, which develops tractor equipment that conducts plant-by-plant analysis for the application of inputs.[30]

The commercialization of agriculture data is another instance of a broader trend where all data has become monetizable for "surveillance capitalism" based business models.[31] The rise of generative AI has also made concerns around asymmetric extraction more acute. There are, therefore, concerns about "data grabs" and the dispossession of farmers from control over a valuable resource.[32] Further, the value chain does not end at the accumulation of the data but often has the objective of employing algorithms that profile and sort populations in consequential but still not fully transparent ways. For instance, there are concerns that data from farmers, even at the aggregate level, can reveal crucial information about the health of soil, nutrient use, cropping patterns, etc. This can be valuable information for input providers, who could drive up prices for crucial inputs for particular farms or particular regions. Data that provides insight into the productive potential of land is also an important indicator of the value of land.[33] Data related to cropping patterns can also provide a crucial advantage in commodities and futures markets.[34] Another example is weather derivatives, which are a climate risk product traded in global financial markets. They cover businesses for departures from typical weather conditions.[35] Payouts are triggered not by actual losses but by the occurrence of specified meteorological conditions. Thus, financial markets can be very data-dependent, with algorithms and models frequently crunching streams of data to make decisions that have significant impacts on societies.

With the global push towards digital agriculture, the dynamics of these data grabs, while they typically originate from the Minority World, are also being replicated in the Majority World. Data grabs in the latter also risk facilitating actual resource and land grabs, and the concentration of land holdings.[36] A 2018 study by GODAN[37] also noted that Open Data by itself risks making farmers even more vulnerable to existing asymmetries of financial, commodity, and information

---

[30] Michal, Lev-Ram. "John Deere Is Paying $305 Million for This Silicon Valley Company." Business news. Fortune.Com, September 7, 2017. https://fortune.com/2017/09/06/john-deere-blue-river-acquisition/.

[31] Some have described it as a new variant of colonialism. See, for instance, Couldry, Nick, and Ulises Ali Mejias. "The Decolonial Turn in Data and Technology Research: What Is at Stake and Where Is It Heading?" Information, Communication & Society 26, no. 4 (March 12, 2023): 786–802. doi:10.1080/1369118X.2021.1986102.

[32] Fraser. "Land Grab/Data Grab:…".

[33] Beer. "Ownership of Open Data:.."

[34] Rasmussen, Neal. "From Precision Agriculture to Market Manipulation: A New Frontier in the Legal Community." Minnesota Journal of Law, Science & Technology 17, no. 1 (February 1, 2016): 489. https://scholarship.law.umn.edu/mjlst/vol17/iss1/9.

[35] Bates, Jo. "Big Data, Open Data and the Climate Risk Market." In Carbon Capitalism and Communication, edited by Benedetta Brevini and Graham Murdock, 83–93. Cham: Springer International Publishing, 2017. doi:10.1007/978-3-319-57876-7_7.

[36] Fraser. "Land Grab/Data Grab…".

[37] Global Open Data for Agriculture and Nutrition, an international initiative promoting Open Data for agriculture and food systems.

flows and can undermine their livelihoods.[38] As a result, data related to rights and claims over resources can be sensitive in such contexts and a source for contestation.[39] As an example, the International Water and Management Institute does not make certain data sets openly available, as exposure of shared resources can generate conflict.[40] When locally-open-data sets are indexed or hosted on a global data platform, they may become more visible and available for use by better-resourced stakeholders. This heightens the imperative to explore ways of democratizing benefits from open-data infrastructures, particularly across geographies.

---

[38] Maru, Ajit, Dan Berne, Jeremy De Beer, Peter Ballantyne, Valeria Pesce, Stephen Kalyesubula, Nicolene Fourie, Chris Addison, Anneliza Collett, and Juanita Chaves. "Digital and Data-Driven Agriculture: Harnessing the Power of Data for Smallholders," 2018. doi:10.7490/F1000RESEARCH.1115402.1.

[39] An FAO Official quoted in Ferris, Lindsay, and Zara Rahman. "Responsible Data in Agriculture," 2017. doi:10.7490/F1000RESEARCH.1114555.1.:"forests will often be considered 'uninhabited' according to official data. This is often inaccurate. At the same time if there is information about the forest inhabitants and they do not have their rights guaranteed, extractive companies may come in and displace these communities. We are seeing more and more cases of extractive industries pushing forest communities out of their traditional lands and resources."

[40] Ferris, Rahman. "Responsible Data in Agriculture."

MITIGATIONS

The Open Data Charter suggests that adopting an "open by default" approach does not imply "automatically" making data available to everyone but an approach that seeks to build trustworthy, fair, and accountable data practices. This includes considerations about enhancing protections when needed, giving voice and agency to people affected by the use of data, and mitigating risks that can cause harm.

While exploring these mitigations, it is important to consider that the severity and likelihood of the risk also depend on the kind of data that is made openly available. Mitigation strategies could be more light-touch in the context of, for instance, atmospheric data but stricter if the data represents (or have consequences for) human behavior in any way. A greater level of scrutiny may be attached to data sets that may have originally contained personally identifiable data, even if they have since been anonymized or aggregated. In particular, additional scrutiny may be warranted if data sets are crowdsourced or opened up expressly for their inclusion onto a platform, particularly in the context of development or climate-related projects in the Majority World.

Many of the challenges and risks that an open-data platform may raise stem from the broader social and political contexts in which they may exist and from where data sets may be indexed. As such, there may be few catch-all solutions that can fully address these challenges, and many may not be within the remit of an open-data platform. Risks may be difficult to anticipate and may be context-specific. Even so, it is worth scrutinizing the different ways open-data infrastructures mediate or amplify these dynamics.

While these concerns related to Open Data are crucial to consider, and mitigations necessary, they need not be seen as an argument against making data openly available per se. Issues such as bias or asymmetric benefit may, in fact, only come to light because data has been made openly available. The approach in this section is, therefore, to explore governance mechanisms and organizational approaches as ongoing practices that can mitigate some of these risks.

Clarity in Vision and Scope

As the Open Data Charter suggests, a focus toward "publishing with a purpose" can be more impactful than simply publishing as much data as possible, as soon as possible. As a starting point, an open-data platform can develop a theory of change that recognizes the limitations of Open Data, treating it as an instrumental and not an intrinsic good.

Considering the broad nature of a category like climate—and agriculture-related data, developing a clear sense of scope around what kind of data sets may be within the platform's

remit can also enable better outcomes. Based on a study of various open-data projects and their impact, the GovLab found that initiatives with a clear target or problem definition have a better impact.[41] This can also prevent function creep, which can also increase the risk profile of data sets indexed and aggregated by the platform.

The Open Data Charter's Open Up Climate Data Guide suggests that understanding the use cases for opening up a data set, along with the capacities and needs of both data providers and prospective users, can be crucial to prioritizing what information to curate and publish. For climate-related data, this could mean early thinking about how particular data sets can inform or improve climate mitigation or adaptation or enhance the ability to measure their results and impact. This approach can ensure that limited resources are used in the most effective ways.

The GovLab's findings also encourage setting up crowdsourced problem inventories, to which users can contribute specific questions and answers, which can help define open-data projects.[42] An example is the UK Ordnance Survey's GeoVation Hub, which poses specific challenges ("How can we improve transport?") for stakeholders to respond to using the Open Data they make available. It may be useful to create 'problem and data definition toolkits' that help formulate clearly defined challenges and connect with the potentially useful data sets.

Usage Monitoring and Key Performance Indicators

Open-data platforms often collect and analyze statistics related to data use on the platform. Usage stats and analytics could be designed to focus specifically on measuring the diversity of stakeholders that use the platform, the kind of data sets that are considered most valuable and impactful, and the nature of uses that data sets are put to.

Generating these analytics (the data from which could also be made openly available for analysis) could provide insight into the extent to which risks related to unequal capacities and asymmetric impact may be manifesting on the platform and can serve as an evidence base for designing responses. Making usage-related data openly available can also enable transparency and third-party, independent audits in support of these objectives.

Objectives related to diverse and meaningful impact can be incorporated into key performance indicators when evaluating the success of the platform. Stakeholder and, particularly, citizen participation in determining these metrics to measure the results of Open Data can also be critical for success. Findings from the GovLab's report on Open Data Impact propose creating a metrics bank of important indicators, with input from stakeholders, researchers, and experts in the field.[43]

---

[41] Verhulst, Stefaan, and Andrew Young. *Global Impact of Open Data*. [Place of publication not identified] O'Reilly Media, 2016.

[42] Verhulst, Young. *Global Impact of Open Data*.

[43] Verhulst, Young. *Global Impact of Open Data*.

Impact Assessments

Building on usage analytics-related practices, a potential strategy to address concerns related to privacy, bias, and ethics is to incorporate impact assessments at various thresholds related to data collection, use, and access. Impact assessments are commonly used in various domains (for instance, environmental regulation and data protection frameworks) to determine the risks that may arise from a particular project, the likelihood and severity of such risks, and to implement (and document) mitigation strategies.

Impact assessments can assess whether aggregated or anonymized data sets might still pose risks related to identifiability or when data, even if non-identifiable, may still raise ethical concerns (for instance, when data related to claims over contested or scarce resources is made open). Impact assessments could also include processes that enable local and subject matter experts to provide insight with respect to context-specific risks.

There can be specific thresholds at which assessments could be introduced, for instance, when new data sets are hosted on a platform, when new data aggregation and analysis layers are incorporated into a platform, or when certain types of data sets are accessed or shared. As risks can arise over time, when more data sets become available and analytic techniques improve, there can be mechanisms in place for periodic assessments. Open-data platforms often have dedicated teams that monitor cybersecurity and privacy related risks, which can be a model to build upon.

Assessment methodologies can draw from existing approaches, such as the Data Protection Impact Assessments, that data controllers are often tasked with carrying out under the European Union's General Data Protection Regulation. These are usually self-assessments, but there is guidance and methodologies that have been created by various regulators (Such as the CNIL in France, and the ICO in the United Kingdom, as well as the European Data Protection Board). Another example of a data ethics assessment framework is the Open Data Institute's Data Ethics Canvas.

Drawing from these models, potential data users could also be tasked with conducting impact assessments when accessing data. While accessing data would not be contingent on the particular outcomes from an assessment, such a requirement can enable early reflection by data users about the use of data and any risks it may pose. These protocols could also be modeled on existing institutional approaches such as Biobanks, or University Institutional Review Boards, which have assessment and documentation protocols in place to ensure ethical collection and use of data. Open-data platforms can also create guidance and model formats for conducting assessments.

Datasheets and Documentation

Data ethics and machine learning practitioners and researchers have recently begun advocating for the creation of "datasheets" whenever data sets are created in order to mitigate risks of bias and to facilitate greater openness and transparency.[44] Different but related approaches have been proposed, such as "model cards," "datasheets," or "dataset nutrition labels." The objective is typically to document the composition, collection processes, motivations, and recommended uses of data sets. Making this information publicly available assists potential users in understanding and assessing whether a specific data set might be a good fit for their purposes.

The platform can incorporate similar documentation processes when data is first hosted on the platform, and/or create guidance that supports the creation of such documentation when data is first collected. Many of these proposals arise from the machine learning community, where the data may be more sensitive than typical open data sets. However, similar processes, with lighter requirements, could enable safer data use on the proposed platform.

As noted, part of the challenge when sharing data is that the risks that may arise can be complex and emergent. Having robust protocols for evidence and documentation, through such assessments, can aid on-going governance and monitoring processes that assess the utility of data sets, the range of stakeholders they benefit, and the nature of risks that tend to arise. Metrics related to impact assessments and documentation processes can further be incorporated into potential KPIs, and represented as open data related to the platform.

Participatory Governance and Representation

In instances where data sets may represent people or important resources or when data is crowdsourced, introducing stakeholder participation can serve as an important mitigation mechanism. Stakeholder participation can also be incorporated into metric identification, impact assessment, and documentation processes. As an example, in an AI project based in Indonesia, the German Development Corporation (GIZ), along with the Participatory Mapping Network (JKPP) and High Carbon Stock Approach (HCSA), collaborate with local and indigenous communities using a process of Free, Prior, and Informed consent. The objective of such a process is to include diverse perspectives, ensure needs and rights, and balance Open Access with safeguarding against unintended consequences.

Recent interest and experimentation around "data stewardship", and institutional approaches such as data commons, data trusts, and data cooperatives can provide inspiration for different models for enabling people to have a greater voice in data-related decisions.[45] For instance,

---

44 Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. "Datasheets for Datasets," 2018. doi:10.48550/ARXIV.1803.09010.

45 European Commission. Joint Research Centre. *Mapping the Landscape of Data Intermediaries: Emerging Models for More Inclusive Data Governance.* LU: Publications Office, 2023. https://data.europa.eu/doi/10.2760/261724.

data trusts can delegate an independent fiduciary to represent the interests of a group of stakeholders,[46] while data cooperatives have been proposed for more bottom-up forms of collective decision-making. The Open Environmental Data Project proposes a range of collaborative governance models for different use cases.

A truly open global data platform would be strengthened by global participation and multistakeholder dialogue, with a specific focus on foregrounding the perspectives of underrepresented stakeholders at its highest levels of governance. The Open for Good Alliance proposes multistakeholder dialogue and governance that engages a range of stakeholders, particularly when data sets may be used to develop AI models. This includes communities and organizations involved in the collection of the data; communities and organizations who aggregate, curate, and vet such data; small, local institutions who nurture and use such data; people who are represented in such data; large international institutions and companies, who use data; the general public. Experiences from related multistakeholder initiatives, like the United Nations' Green Climate Fund or the Extractive Industries Transparency Initiative.

The impacts of climate change are also projected to be asymmetrical, with different communities being vulnerable to unequal degrees. In some cases, communities that are most vulnerable to the impacts of the climate crisis may also be those that have unequal capabilities with respect to Open Data use. Ensuring diversity and representation in the platform's governance structures can enable long-term sustainability of the platform in the spirit of Public Digital Infrastructure.

Funding and Capacity-Building

There is growing acceptance among open-data advocates that making data available should not be seen as an end in itself but an intermediate step toward actual outcomes. The push for Open Data shouldn't be understood as an alternative to funding services, interventions, and infrastructure.[47] Solutions and resources should also be targeted at empowering stakeholders to benefit from the data and supporting climate action. Analysis and aggregation layers built on top of the data sets can serve as a valuable step towards democratizing use and benefit. However, closer attention can be paid to all types of capabilities that can support (and equalize) the impact of Open Data.

---

[46] For a specific exploration of data trusts for climate impact see: "Enabling Data Sharing for Social Benefit Through Data Trusts: Data Trusts in Climate", https://gpai.ai/projects/data-governance/data-trusts-in-climate-interim-report.pdf (Accessed May 27, 2024).

[47] Bates, Jo. "The Strategic Importance of Information Policy for the Contemporary Neoliberal State: The Case of Open Government Data in the United Kingdom." *Government Information Quarterly 31*, no. 3 (July 2014): 388–95. doi:10.1016/j.giq.2014.02.009.

Asymmetries in benefits from Open Data could be mitigated by exploring and making available resources that support a wider range of stakeholders to use data. This is particularly important when data can easily serve as a raw material for commercial products and surveillance capitalism-based extractive enterprises at the expense of stakeholders that generated the data or are represented in it. These concerns are particularly urgent when considering asymmetries in capabilities and resources across geographies.

The World Resources Institute's study on Open Data Strategies for Climate Action found that fostering networks and communities of practice around Open Data use can be an important step.[48] Many open-data initiatives have also begun committing resources towards training and capacity-building. For instance, GODAN produces a range of tools and resources that support data reuse. Open-data initiatives can also directly provide funding for data reuse. An example is the Global Forest Watch, which runs a small grants and fellowship program. The grants enable a broader range of stakeholders, particularly civil society organizations and individuals, to effectively use GFW data in their advocacy, research, and fieldwork. Providing funding directly may have the benefit of allowing stakeholders and communities to determine for themselves the nature of capacity-building and resources that may be most useful to them in reusing data.

Fostering Intermediaries

Much of the writing on Open Data highlights the importance of intermediary organizations in translating data into meaningful knowledge and benefits for stakeholders.[49] Intermediaries bring together various forms of resources and capital, increasing the utility of data and democratizing the benefits from it.[50] As a result, intermediaries have been described as a keystone species of Open Data ecosystems. In particular, intermediaries can be a way to translate climate data for a wider audience, and to engage groups most directly affected by climate change.[51] In its findings on Open Data Impact, the GovLab therefore suggests supporting intermediaries, particularly those from civil society and in contexts with lower technical capacities.[52]

---

[48] Grinspan, Delfina, and Jesse Worker. "Implementing Open Data Strategies for Climate Action: Suggestions And Lessons Learned for Government and Civil Society Stakeholders." *World Resources Institute*, 2021. doi:10.46830/wriwp.19.00093.

[49] Magalhaes, Gustavo, Catarina Roseira, and Sharon Strover. "Open Government Data Intermediaries: A Terminology Framework." In *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, 330–33. Seoul Republic of Korea: ACM, 2013. doi:10.1145/2591888.2591947. Van Schalkwyk, François, Michelle Willmers, and Maurice McNaughton. "Viscous Open Data: The Roles of Intermediaries in an Open Data Ecosystem." *Information Technology for Development 22,* no. sup1 (August 30, 2016): 68–83. doi:10.1080/02681102.2015.1081868.

[50] Schrock, Andrew, and Gwen Shaffer. "Data Ideologies of an Interested Public: A Study of Grassroots Open Government Data Intermediaries." *Big Data & Society 4,* no. 1 (June 2017): 205395171769075. doi:10.1177/2053951717690750.

[51] Grinspan, Worker. "Implementing Open Data Strategies for Climate Action:.."

[52] Verhulst, Young. *Global Impact of Open Data.*

Open-data platforms can, therefore proactively engage with intermediaries, as focal points for capacity-building and funding programs. Open-data intermediaries can themselves also draw from data stewardship approaches, to implement tailored and trustworthy governance mechanisms with respect to the data they access and use.

SUMMARY OF RECOMMENDATIONS

Clarity in Vision and Scope

Recognize Open Data as an instrumental and not an intrinsic good. Develop a theory of change that acknowledges the limitations of Open Data and proactively attempts to address them. Articulate a clear sense of scope for the platform and the data sets it will host.

Usage Monitoring and Key Performance Indicators

Monitor Open Data usage with a view towards promoting democratized and diverse usage of the platform. Incorporate these objectives in the ways success is measured on the platform. Enable stakeholders to feed into the processes of metric development.

Impact Assessments

Introduce assessment processes to determine risks related to data that is indexed and accessed. Conduct periodic assessments related to ethics, privacy and data protection. Encourage data users to conduct assessments, and support them by publishing internal guidance.

Datasheets and Documentation

Foster transparency and accountability by incorporating practices that document the composition, collection processes, motivations, and recommended uses of data sets. Create an evidence base that can inform monitoring and interventions related to biases and risks.

Participatory Governance and Representation

Enable participation from the data set level up to the platform. Allow stakeholders to participate in metric development, impact assessments, documentation processes, and high-level governance. Pay particular attention to the global asymmetries in data capabilities and climate impacts. Ensure representation in governance structures. Explore data stewardship approaches as models that support participatory governance.

Funding and Capacity-Building

Explore pathways that enable a diverse range of stakeholders to use and benefit from data. Create resources and tools that support data use. Commit funding towards capacity-building.

Fostering Intermediaries

Support intermediary organizations that can translate data access into meaningful benefits. Focus and channel resources and capacity-building efforts through organizations that are rooted in and directly engage stakeholder communities. Explore data stewardship models that enable trustworthy governance with respect to data.

ABOUT OPEN FUTURE

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge and culture.

This study was commissioned by Open Future to explore how the Paradox of Open manifests itself in the context of open-climate data and how climate data can be governed as Digital Commons.

Aditya Singh is a lawyer, researcher, and consultant with a focus on data protection and AI ethics. His current research explores models for collective governance over data in food systems, with the objective of mitigating power asymmetries in the sector. He is also interested in critical and decolonial approaches to data and knowledge infrastructures.

ACKNOWLEDGMENTS

The author would like to thank Nadine Dammaschk for her thoughtful comments and feedback on the report.