Mechanisms to Govern Responsible Sharing of Open Data: A Progress Report

This manuscript (permalink) was automatically generated from Sage-Bionetworks/governanceGreenPaper@76e4232 on June 16, 2020.

Authors

- **Lara Mangravite**
  0000-0001-7841-3612
  Sage Bionetworks

- **Avery Sen**

  Sentripetal

- **John T. Wilbanks**
  0000-0002-4510-0385 · wilbanks
  Sage Bionetworks

Abstract

This report provides a landscape analysis of of models of governance for open data sharing based on our observations in the biomedical sciences. We offer an overview of those observations and show areas where we think this work can expand to supply further support for open data sharing outside the sciences.

The central argument of this paper is that the "right" system of governance is determined by first understanding the nature of the collaborative activities intended. These activities map to types of **governance structures**, which in turn can be built out of standardized parts — what we call **governance design patterns**. In this way, governance for data science can be easy to build, follow key laws and ethics regimes, and enable innovative models of collaboration. We provide an initial survey of structures and design patterns, as well as examples of how we leverage this approach to rapidly build out ethics-centered governance in biomedical research.

As this paper itself will be deposited in GitHub, we also envision a contributory process whereby this inventory can be extended with more resources and links over time. We can envision communities using these design resources to create clearly governed networks. We can also imagine small private collectives amongst corporations and their partners, both academic and smaller businesses, using these designs as a "stack" to govern data science beyond biomedical research.

While there is no one-size-fits-all solution, we argue for learning from ongoing data science collaborations and building on from existing standards and tools. And in so doing, we argue for data governance as a discipline worthy of expertise, attention, standards, and innovation.

Introduction to Data Governance

We have access to more data than ever before. We have access to faster computing resources than ever before. But technology, by itself, will not cures or create the solutions to climate change. The 2010s showed that unrestricted personal data surveillance can hack at the roots of society itself — changing elections and undermining collective trust in truth and experts [1]. Data companies are now struggling to institute governance after-the-fact, building advisory boards, and attaching flags to posts.

Beyond social media and venture capital, however, lies a vast, rich, and often under-explored field of data governance. Particularly in the sciences, it is not controversial to believe that creating technology requires also creating the practices that govern its use. These scientific cultures often already work under regulation, and thus their existing practices predict life under data protection regulations and privacy laws.

Data scientists in the 21st century, like physicists in the last one, must reckon with the power they have to change the world forever. Data about people and the world is inseparable from the values that underlie its creation and application. But data scientists, particularly in non-regulated settings, often work without institutions, guideposts, or guardrails — without governance. The outcome is a mishmash of algorithmic inference that is inaccurate at best, racist and divisive at worst [2].

Governance is a broad term. In a general sense, governance is a system of setting policy to encourage and prohibit behaviors, to monitor and enforce such policy, and to issue penalties or rewards accordingly. Governance is embodied in laws or other rules to tell humans (or machines) what they cannot do, or to incentivize them to try to do things they might not do otherwise. It can also be hardwired into tools (e.g. choice of data type) to prevent them from ever being able to do prohibited things, or to only be able to do things that are encouraged. Some aspects of governance are explicit — either codified into rules (written in natural language or computer code) or embodied in the very structure of equipment. Other aspects of governance are implicit — existing as cultural norms or tacit knowledge passed from person to person in practice.

In the context of biomedical research, we define governance as **the freedoms, constraints, and incentives that determine how two or more parties manage the ingress, storage, analysis, and egress of data, tools, methods, and knowledge amongst themselves and with others**.

Each step requires software, storage, compute power, know-how, and access to external digital resources. Each step further involves communication and negotiation. The resulting co-created knowledge is also more than just a publication in an academic journal, comprising mathematical models, networks, graphs, or other complex analyses, systems, or representations. Validating the claims in that knowledge may require access to analytic scripts written expressly for the data as well as enough infrastructure to re-run the entire analysis. The “data” in data science thus means more than just a literal data file.

Complicating matters further, different scientific communities manage governance very differently — high energy physics is the ultimate collaborative and standards-based field for governance, while in the field of biology governance remains individualized and often artisanal. This difference in governance intertwines with the nature of data and its collection: physicists are long accustomed to sharing large-scale equipment to generate, while biologists typically work in individualized laboratory settings with local data-generating equipment [3].

This paper argues that the “right” system of governance (including the “right” types and quantities of resources for governing) is determined by the nature of the collaborative activities intended. Concordantly, if our current system of governance is misaligned with our collaborative intent, and we want to transition to a system that is better aligned, then our transition strategy must be determined by an understanding of both the "as-is" and the "to-be" forms of collaboration.

As we will see in greater detail below, much scientific data governance revolves around how ***available*** data will be (i.e., how many and what types of people can access it) and how many ***freedoms*** are given to those who can access it (i.e., what conditions limit how can they use it). To understand these attributes of collaboration in context, consider the following examples.

First, a sensitive data set composed of a million mammograms with identifiable information could be powerful for studying breast cancer, if made widely available. But such a data set is enormous, costing tens of thousands of dollars in cloud fees just to transfer and store it, and more to analyze it. Further, its privacy implications would make its distribution complex, significantly constricting users. The right data governance approach can unlock that data set while balancing cost and privacy, allowing rigorously vetted users to apply for access to compute on it, privately, while preventing data extraction [4] and undesired uses [5].

Similarly, a data set that extracts sensor information about Parkinson's disease from the phones of 15,000 people could drive new insights both into how the disease progresses, and into how we can use phones to study disease. We can study memory, phonation, and gait over time with no devices other than a smartphone. Such a data set would be far less identifiable than the mammograms, and smaller. Thus, its data governance can, in turn, leverage broader distribution and lower barriers to access, allowing a large audience of lightly vetted users to download, process, and publish new insights [6].

These two examples are real-life examples from Sage Bionetworks (the DREAM digital mammography challenge and the mPower observational study of Parkinsons). They connect fundamentally through the conviction that there is not a simple answer to how we might govern data. Instead, we must look to technical realities, contractual methods, and, most importantly, how they can work together to unlock responsible data use in the service of more reliable claims from data science.

Importantly, our own focus is in biomedical research data sharing. This focus invokes a series of privacy laws and regulations relating to the identifiability of the data — its mismanagement has done direct harm to people [7] and the variable nature of the data makes aggregation from across multiple sources complex. For us, the context under which data were collected will bound the ways the data can be meaningfully used. Thus, data sharing in our space regularly requires nuance and customization [8]. While our space has been an outlier compared to consumer data and government open data, the increasing likelihood of data protection legislation around the globe means our examples may reflect a looming future for all data about individuals.

The good news is that these examples also point to a designable future that leverages a few key governance models to capture most of the common governance patterns encountered, with customization around the edges to reach the most complex patterns specific to any particular data set. We introduce the concept of **governance structures** to describe the models that we see forming over time. We also introduce the concept of **governance design patterns** to describe reusable elements of governance, such as standard contract language or user interfaces. These design patterns offer another path to governing open science: they codify a variety of research collaborations which, in turn, make more data more open. They represent a potential future direction for data governance as a discipline.

The rest of the paper will be as follows. We will account for major structures of scientific collaboration, including how suited each structure is to meeting open science goals. We will detail each governance structure in terms of a few key attributes, including availability and freedom. Next, we will discuss the governance design patterns associated with each of the collaborative patterns. This will lead into a discussion about how to transition data resources from one governance design to another — a form of data publication. Finally, we offer two example projects currently underway at Sage Bionetworks to make concrete the preceding concepts of governance structures and governance design patterns.

Key Governance Concepts

Collaborations can embody a mixture of different forms of data movement and exportation, freedoms and restrictions on analysis, and openness and closure. In practice, we find that parties hold different ideas of what open, closed, accessible, findable, and interoperable actually mean. We are inspired by the FAIR principles — meaning an analysis of data as Findable, Accessible, Interoperable, Reusable — that are widely adopted in the life sciences (Wilkinson et al, 2016), but their strength as a general standard also limits their application inside our own governance requirements. In this section, we will explain our own definitions of these and other terms.

We have chosen Availability and Freedoms as our two key attributes for characterizing collaborations, recognizing that both terms contain multitudes and, in some cases, may conflict with existing uses.

> Availability: the size of the population to whom a specific data set is available.

Our first attribute is the **total potential size of the user base** for a given data set. This allows us to segment various kinds of collaboration projects by a clear metric of user population size. Here we examine concepts such as: exclusion of classes of users based on their employment (e.g., non-commercial restrictions) or training (e.g., allowance for formally trained and federally funded scientists), and requirements for transaction costs, such as registration, statements of intended use, and identity validation. Community scientists from non-traditional research settings are often excluded by these classes when concepts of availability are not intentionally designed into governance structures.

> Freedoms: the scale of the constraints under which a user must work.

Thus, our second attribute concerns the ways that governance **grants or restricts users' rights to use the data**. This measure is more qualitative than user population size, and forms more of a heuristic than a metric. Questions addressed include: Can the data be downloaded or redistributed? Are there fields of inquiry that are excluded? Is ethical oversight needed? Is exploratory use granted? And so forth.

> Open: the data are available under an explicit, pre-negotiated license that guarantees, at minimum, their ability to be downloaded and used for unrestricted data analysis.

This definition focuses on granting, in advance, the rights to reuse data locally, and includes open licenses such as Creative Commons Zero, the Open Database License, and Microsoft's Open Use of Data Agreement.

> Closed: the data are only available via petition to the data owner.

This definition encompasses most day-to-day data practice.

> Restricted: there are regulatory or ethical restrictions on the data that necessarily constrain their availability.

This definition contemplates, in particular, privacy and data protection law, but also constraints such as those present on data from indigenous peoples and tribal nations.

> Governance structure: a form of research collaboration governance, as characterized by the number and nature of relationships among parties, the relative degrees of availability of data and freedoms to use them, and instantiated by a contract (or other legal language) plus other, subordinate modules (e.g., user qualification mechanisms).

This definition asserts that governance takes forms over time that are observable and repeatable.

> Governance Design Pattern (or simply “pattern”): a module within a governance structure, i.e., a concrete artifact (e.g., boilerplate language for rules, or algorithms to data monitor sharing).

This definition asserts that governance structures are composed of specific and reusable patterns.

Governance Structures

Effective data use traditionally starts not with, “what data do we have?” but instead with, “what hypothesis do we propose to explore?” The advent of machine learning adds to this, “what hypotheses are afforded by these data?” These two questions create specific local cases with wide and diverse requirements: what data to acquire, how to bring them into systems, how to store them, how to analyze them, how to share downstream knowledge.

Because of this diversity, what is appropriate data handling in one project may not be appropriate in another. Notably, open approaches to data access do not consistently lead to their responsible and reliable reuse [9]. This has made the process of data governance itself difficult to “open source” — simply giving away a clinical protocol, data management process, or metadata schema under an open source license does not ensure meaningful data reuse.

Openness itself merits interrogation. A key requirement for data governance to consider in this context is: **what does “open” mean to different people, and different groups, over time**? Openness does not work for everyone, everywhere. Many groups traditionally under-represented in medical research have endured systemic, ongoing betrayals of trust in how data and samples are collected and used (Bardill and Garrison, 2015). These same groups are also often the least served by the medical systems, and the most surveilled by the state [10]. We must therefore design governance structures and governance design patterns that support the contextual desire to restrict data availability [11]. We must also develop the ability bring every group into governance designs, to explore when, how, and to whom they might be comfortable allowing access over time, and governance appropriate for those uses (Wilkins, 2020).

While research collaborations are unique, there are commonalities they share. At Sage Bionetworks, we have observed regular, relatively stable structures that work well over time. We will refer to these as **governance structures**.

The major governance structures we have observed are the following:

- **Pairwise** (One-to-one): Two parties agree to work together and/or share on a data set in some fashion, typically with a closed contract or an informal agreement. The negotiation terms depend on the relative status of the parties and/or the value of the data and knowledge.
- **Open source** (One-to-many or some-to-many): data are distributed for reuse with a license defining reuse rights and conditions. The creator is in charge of the negotiation at first (choice of license), but then rights to analyze and redistribute are permanently transferred to the user. This is typical of a centralized project in the sciences, i.e., the Human Genome Project (Collins et al., 2003).
- **Federated Query** (Many-to-many, via platform): Data are housed in a variety of locations, and users are able to query to those local data simultaneously. Typically restricted to pre-configured queries (rather than data exploration) and may require registration before use (Schwarte et all, 2011).
- **Trusted research environment** (Many-to-some): data are housed in a central location under a contractual regime including data transfer and use agreements. Users apply to use the data. Users must “visit” the data rather than download them, agree to be known, and, in some cases, agree to be surveilled by a data steward (Lane and Shipp, 2007).
- **Model-to-data** (One-to-many): Data are held by a steward who is responsible for running algorithms on the behalf of researchers. In some cases, a synthetic version of the data may be released openly to facilitate model training. Researchers develop algorithms, send them to the steward, and receive back output of

their analysis as run on the real dataset. The variety of analyses that may be performed is restricted by this structure, because the data steward must ensure data are specifically curated for any analytical question at hand. (Guinney and Saez-Rodriguez, 2018),

- **Open citizen science** (Many-to-many): Rights to use and distribute data are often fully decentralized via license or contract. Open citizen science is a peer-to-peer version of open source science. (Greshake et al, 2019)
- **Clubs and Trusts** (Some-to-some): This is a common version of collaboration in biomedical research. Clubs (Ostrom, 2010) and Trusts (McDonald, 2019) are versions of a common pool resource: a group of people and/or institutions who agree to share resources towards a common goal. Control over the development and negotiation of data sharing and use terms is often held by the founders / settlers (and/or funders) and then can be distributed amongst club participants (Ostrom, 2010). Importantly, **clubs that operate in the cloud can easily publish data products that are more "open" than the club itself**.
- **Closed**: Data are held privately by a single party.
- **Closed and Restricted**: Data are held privately in order to protect a population, meet a legal requirement, or protect a secret.

Table 1: Governance structures and their attributes

Governance structure	Number and linkage of parties	Degree of data Availability	Degree of freedom to use data	Challenges common to the governance success	Primary governance design pattern
Pairwise	One-to-one	Medium/High	Medium/High	Uneven status of parties, value of data	Informal or closed contract
Open Source	One/some-to-many	High	High	Rights permanently granted to user	License
Federated Query	Many-to-many, via platform	High	Medium/Low	Defection of creators	Contract and club rules
Trusted Research Environment	One/some-to-many	Medium/Low	Medium/Low	Users agree to be known, surveilled	Data transfer and use agreements
Model-to-Data	One-to-many	High	Low	Not all who apply can use data	Restricted analyses, data curation
Open Citizen Science	Many-to-many	High	High	Capacity for analysis is uneven	Contract or license
Clubs, Trusts	Some-to-some	Medium/Low	High	Easy to create things governed more liberally. Trusteeship can be revoked.	Club / Trust rules
Closed	Many (to none)	Low	High	Fundamental limits to collaboration	Public laws, security protocols
Closed and Restricted	Some (to none)	Low	Low	Fundamental limits to collaboration	Public laws, security protocols

How many parties are involved, who is in charge of the negotiation, how significantly are the data regulated or the sensitive nature of the data — these are all attributes that can be used to characterize the governance structure. From the many potential attributes, we choose to organize governance structures by a) how broadly available they are — the total potential size of the user base — and b) how many freedoms a user has to freely use and distribute the data.

These two attributes are “upstream” from decisions such as which license to use, and form a pair of axes on which to orient structural governance analysis. These attributes provide a standardized form to describe governance structures in which one, some, or many parties either provide or use data, and the freedom to use and distribute data is variable.

References

1. **Antisocial media: how facebook disconnects US and undermines democracy**
Siva Vaidhyanathan
Oxford University Press (2018)
ISBN: 9780190841171

2. (2018-07-16) http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

3. **Epistemic cultures: how the sciences make knowledge**
K. Knorr-Cetina
Harvard University Press (1999)
ISBN: 9780674258938

4. **Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms**
Thomas Schaffter, Diana S. M. Buist, Christoph I. Lee, Yaroslav Nikulin, Dezso Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, … and the DM DREAM Consortium
JAMA Network Open (2020-03-02) https://doi.org/gg2dvg
DOI: 10.1001/jamanetworkopen.2020.0265 · PMID: 32119094 · PMCID: PMC7052735

5. **Alternative models for sharing confidential biomedical data**
Justin Guinney, Julio Saez-Rodriguez
Nature Biotechnology (2018-05-09) https://doi.org/gg2dvh
DOI: 10.1038/nbt.4128 · PMID: 29734317

6. **The mPower study, Parkinson disease mobile data collected using ResearchKit**
Brian M. Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E. Ray Dorsey, … Andrew D. Trister
Scientific Data (2016-03-03) https://doi.org/gg2dvj
DOI: 10.1038/sdata.2016.11 · PMID: 26938265 · PMCID: PMC4776701

7. **Genomic Justice for Native Americans**
Nanibaa’ A. Garrison
Science, Technology, & Human Values (2012-12-21) https://doi.org/gg2dvk
DOI: 10.1177/0162243912470009 · PMID: 28216801 · PMCID: PMC5310710

8. **Perceived Benefits, Harms, and Views About How to Share Data Responsibly**
Phaik Yeong Cheah, Decha Tangseefa, Aimatcha Somsaman, Tri Chunsuttiwat, François Nosten, Nicholas P. J. Day, Susan Bull, Michael Parker
Journal of Empirical Research on Human Research Ethics (2015-08-21) https://doi.org/f7pmw8
DOI: 10.1177/1556264615592388 · PMID: 26297749 · PMCID: PMC4547202

9. **On the Reuse of Scientific Data**
Irene V. Pasquetto, Bernadette M. Randles, Christine L. Borgman
Data Science Journal (2017-03-22) https://doi.org/gf87pf
DOI: 10.5334/dsj-2017-008

10. (2018-02) https://datasociety.net/wp-content/uploads/2018/02/DataSociety_Fairness_In_Precision_Medicine_Feb2018.pdf

11. **Ethical aspects of data sharing and research participant protections.**
Michael W. Ross, Martin Y. Iguchi, Sangeeta Panicker
American Psychologist (2018-02) https://doi.org/gc4823
DOI: 10.1037/amp0000240 · PMID: 29481107