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

Authors

- **Lara M. Mangravite**
  [0000-0001-7841-3612](#)
  Sage Bionetworks

- **Avery Sen**
  [0000-0002-1303-7449](#)
  Sentripetal

- **John T. Wilbanks**
  [0000-0002-4510-0385](#) · [wilbanks](#)
  Sage Bionetworks

- **Sage Bionetworks Team** [†]

[†] **Sage Bionetworks Team**: Tori Allen, Vanessa Barone, Brian M. Bot, Megan Doerr, Shira Grayson, Justin Guinney, Abigail Karikari, Emily Lang, Lea Martin, Sarah Moore, Ann Novakowski, Allie Seroussi, Christine Suver

Abstract

This report provides a landscape analysis 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

Data scientists in the 21st century, like physicists in the last one, must reckon with the power they have to change the world forever. 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.

In medical research, we are regulated towards values in data connected to beneficence, for example – research should help, not make things worse, for a patient. In consumer technology the values discussion is often around free speech, or growth. Either way, data about people and the world is inseparable from the values that underlie its creation and application.

It is not just data points, but data collections, which enable and constrain the world of possible analysis, that connects to those values. 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,[9] 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.[10] 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.[11] These same groups are also often the least served by the medical systems, and the most surveilled by the state.[12] We must therefore design governance structures and governance design patterns that support the contextual desire to restrict data availability[13] while recognizing that frameworks exist for investigating boundaries and information sharing.[14] 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.[15]

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.[16]
- **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.[17]
- **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.[18]
- **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.[5,19]

- **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.[20]
- **Clubs and Trusts** (Some-to-some): This is a common version of collaboration in biomedical research. Clubs[21] and Trusts[22] 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.[21] 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.
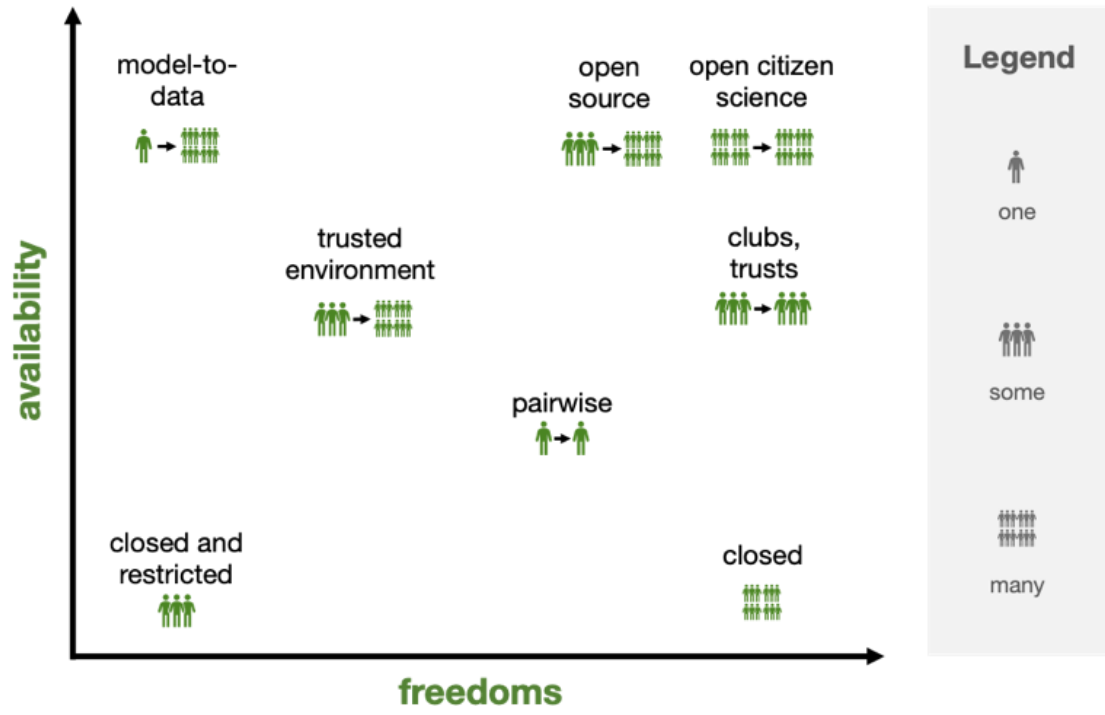
Figure 1: Governance structures and their relationship types, by relative amounts of availability and freedom

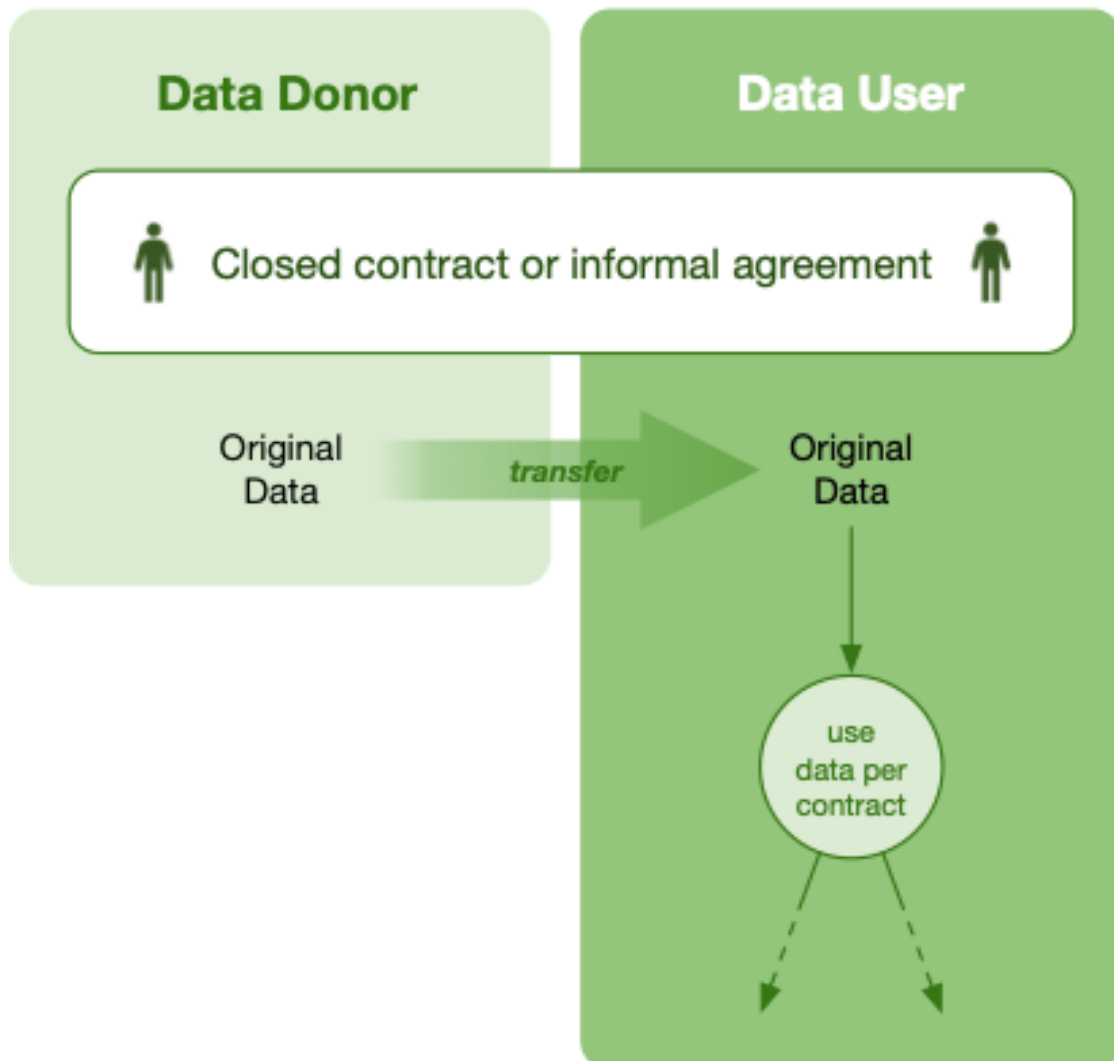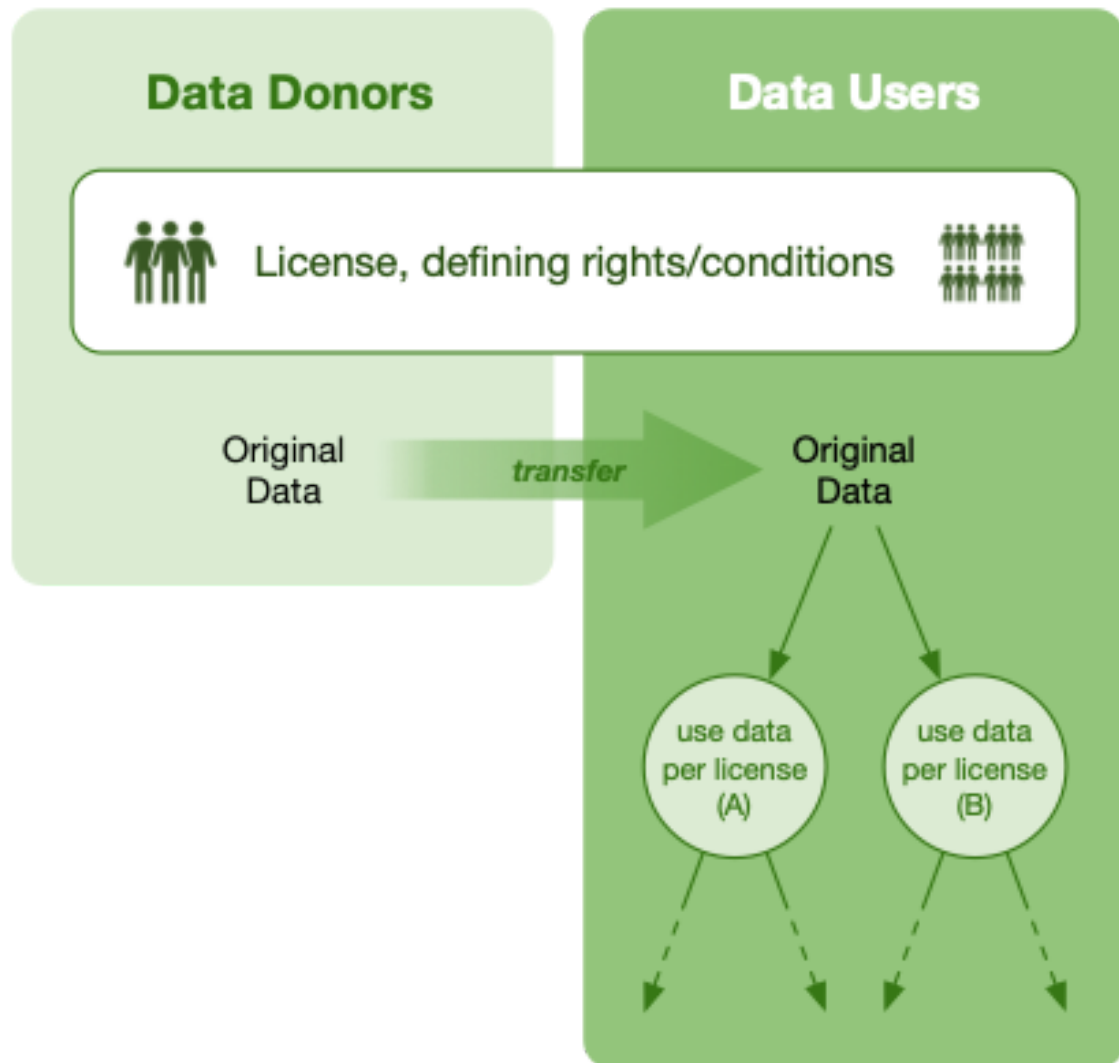**Figure 2:** Governance structure and process flow for pairwise collaborations

Figure 3: Governance structure and process flow for open source collaborations
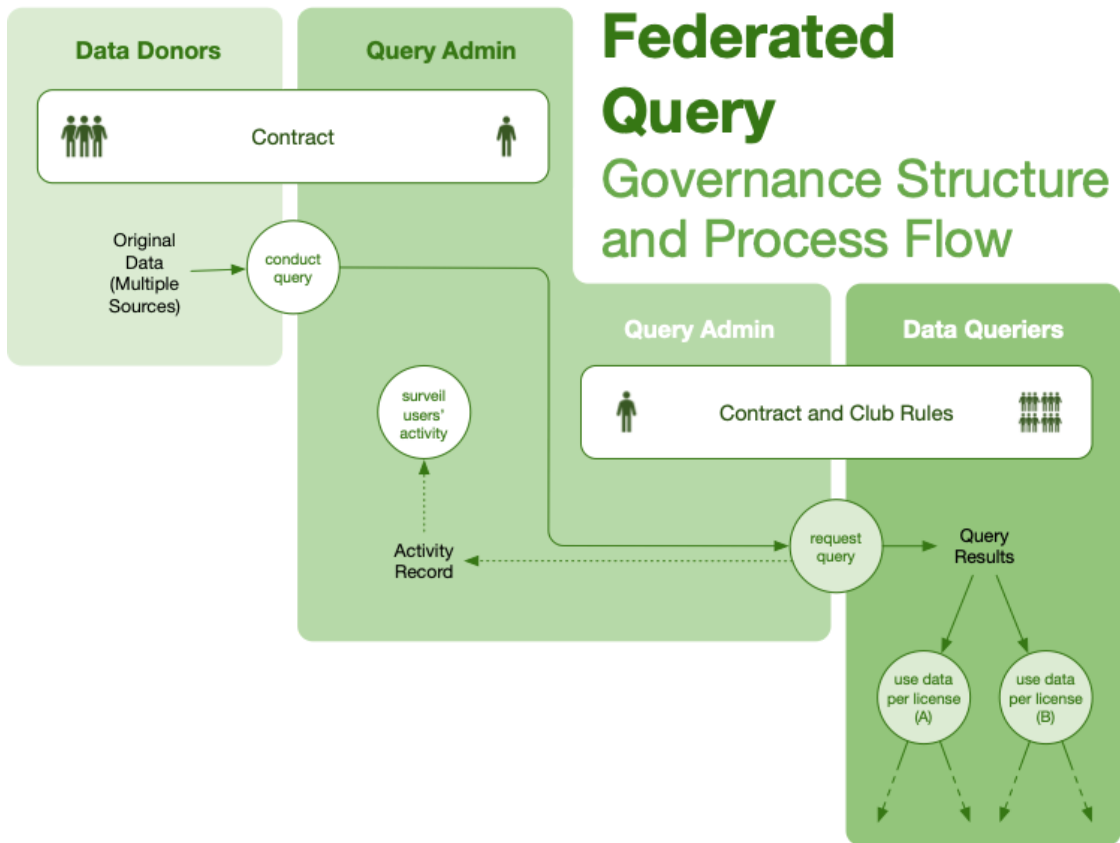
Figure 4: Governance structure and process flow for federated query
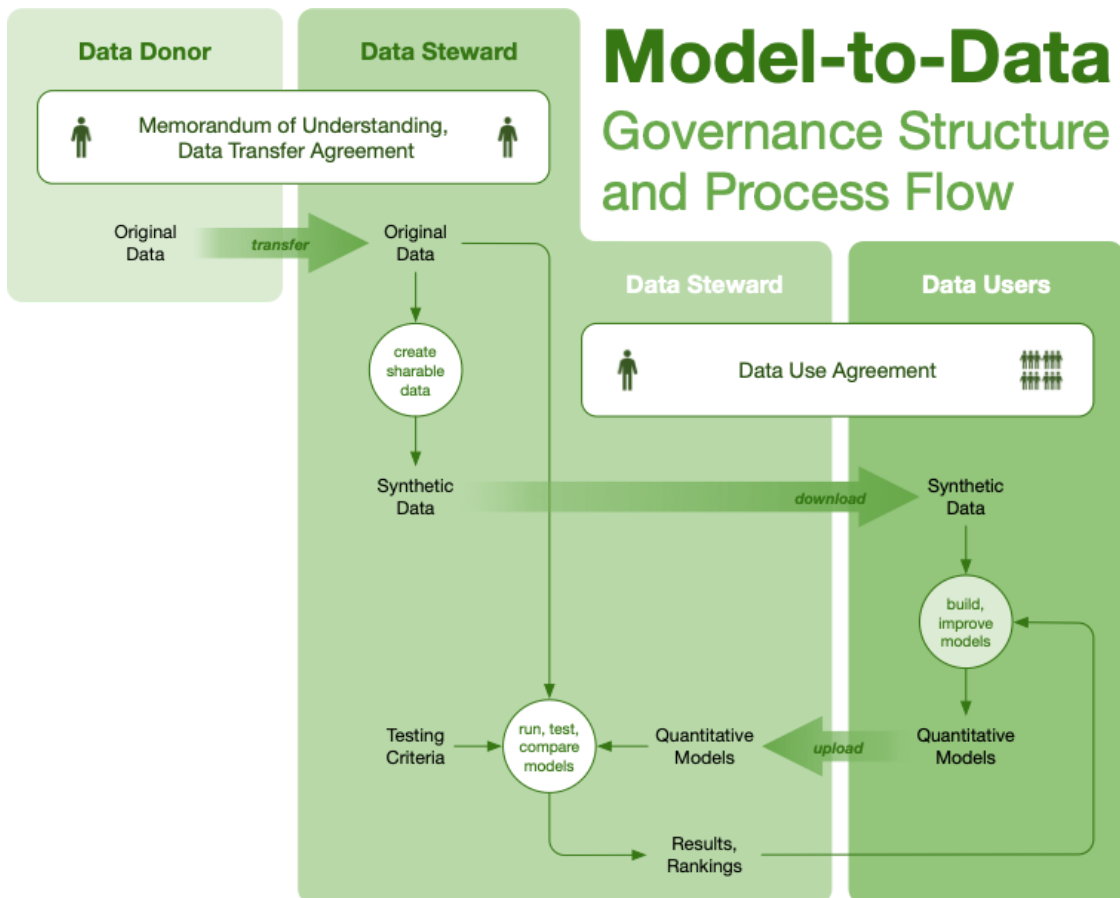

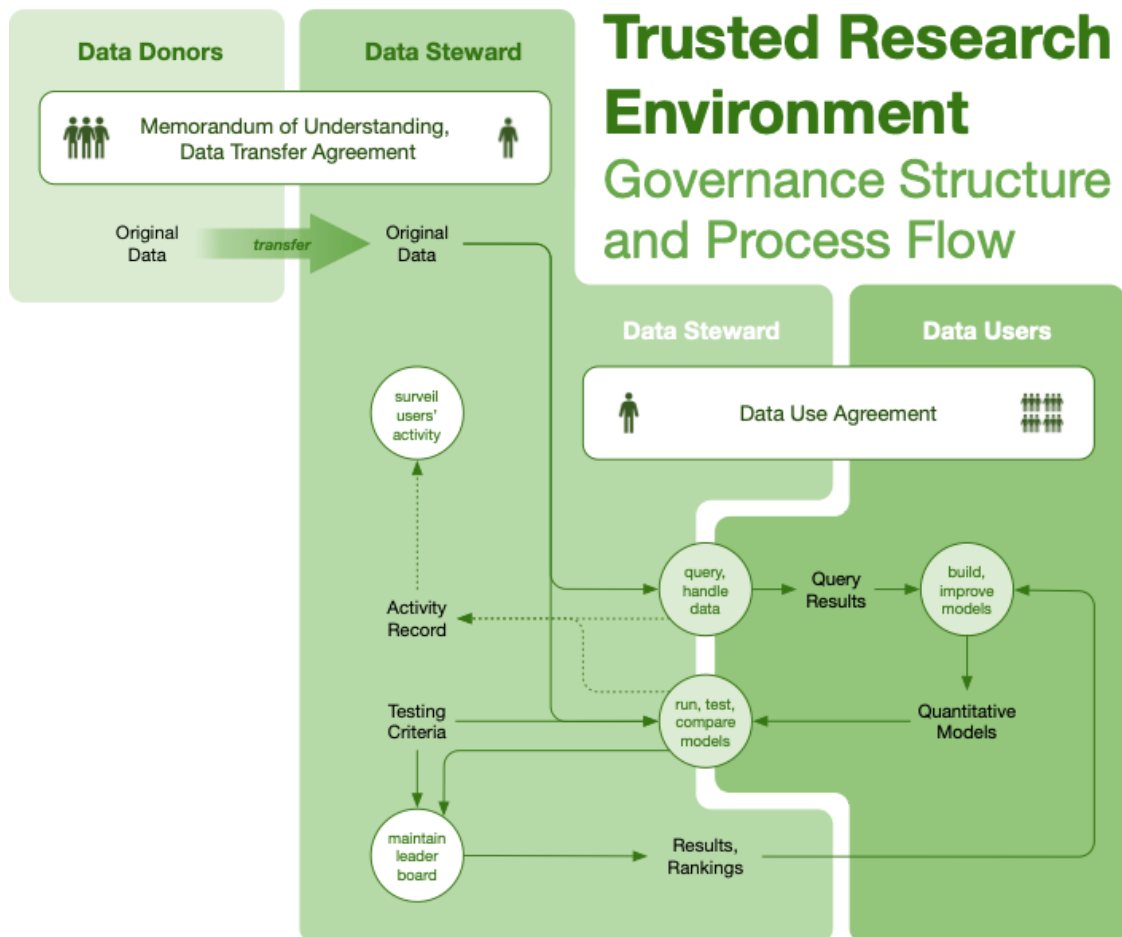**Figure 5:** Governance structure and process flow for model-to-data

Figure 6: Governance structure and process flow for trusted researcer environment

Governance structures are abstractions, not turnkey solutions that we can provide to would-be collaborators, whether open or closed. These structures are most useful to orient data collaborations: do we want a club, or a trusted environment? By asking these questions about how available data will be, and how much freedom a user will have to analyze and distribute, we can quickly identify the goals of the collaborating parties and move into drafting contract structures.

These structures employ governance design patterns: standard contracts and contractual language, user interfaces, teaching toolkits, and software templates. **These patterns can be assembled into a governance structure faster than drafting from scratch, while leveraging legal rigor of past work for regulated or protected data collaboration.**
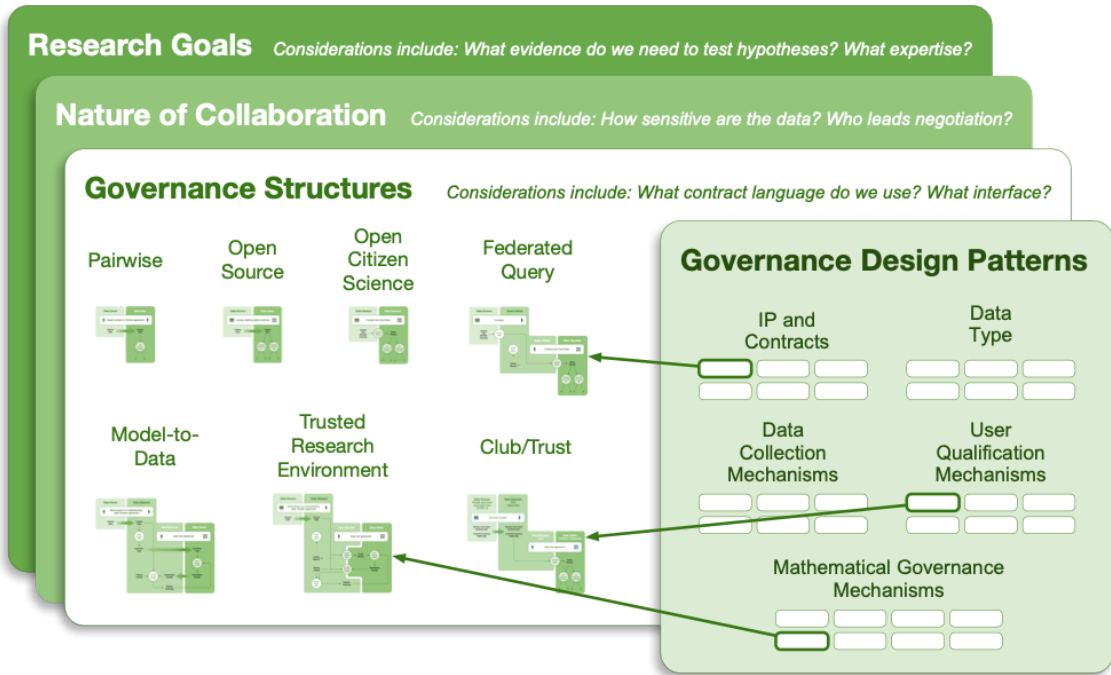
Figure 7: Decisions about governance should start with research goals and nature of collaboration, then move to what structure and design elements are needed.

Governance Design Patterns

Like templates for building a spreadsheet or website, **governance design patterns are generalizable solutions for common challenges and decision points in collaboration**.

Governance design patterns are often not finished products that can be transformed directly into contract. Any real-life governance structure will combine nuances about specific data transfer and use with existing design patterns such as standards, toolkits, reference implementations, process knowledge, and pedagogy. Governance design patterns themselves must often be customized, or "filled out," from their default forms. Authors of these design patterns can build maps of what kinds of governance design patterns work well with each other, which don't, and which, in turn, serve as inputs into new design patterns.

Types of Governance Design Patterns

Data collaborations in biomedical research often feature scientists negotiating legal language directly. Their selection, combination, and application of legal terms is often ad-hoc, often leading to missing language, conflicting language, or misinterpretation. Standardizing contract language patterns and defining their interactions creates a common language for governance early in collaborations. And there is precedent, as the CORE platform hosts a large and growing library of reusable resources[23] as design patterns.

More importantly, standard, yet customizable, reusable language creates multiple benefits. Such language is not biased to anyone in the negotiation, meets legal and ethical norms and requirements often missed by unskilled negotiators, and supports descriptive annotations to facilitate understanding. Standard language also supports automated data governance regimes such as machine review of data use agreements.[24]

Therefore, **the primary governance design pattern at Sage Bionetworks is legal language, usually (but not exclusively) situated inside a contract or a license**. Other patterns include teaching toolkits and user interfaces.

There is a wide range of promising areas to explore for establishing standard governance designs and their constituent design patterns. We review these areas here.

Licenses and Agreements

Licenses, contractual agreements, and regulatory documents can be standard documents, standard paragraphs to be inserted into custom documents, or templates with significant content filled in by the users.

Table 2: Licenses and agreements

Design Pattern	Description	Example
Open license	A license or contract which contains provisions that allow other individuals to reuse data with specific freedoms and conditions. Connected to intellectual property laws.	Creative Commons,[25] Open Database License,[26] Microsoft Open Use of Data Agreement[27]
Data use agreement	An agreement that allows use of restricted data in a specific way. Connected to health privacy laws.	Health Care Systems Network Data Use Agreement Toolkit[28]

Design Pattern	Description	Example
Data transfer agreement	An agreement that allows transfer of data from one party to another not associated with the first party under certain terms and conditions. Connected to data protection laws.	European Commission Standard Contractual Clauses,[29] Uniform Biological Materials Transfer Agreement[30]
Data processing addendum	Standard contractual clauses for transfer between data controllers and data processors. Relevant in GDPR transfers as well as emerging state legislation in the US.[31]	IAPP Sample Addendum Addressing Article 28 GDPR and Incorporating Standard Contractual Clauses for Controller to Processor Transfers of Personal Data[32]
Public domain	A declaration that no intellectual property rights exist in a data set or database.	Creative Commons CC0,[33] Public Domain Database License[34]

Data Collection and Ingest Mechanisms

Data has to be ingested into systems, so that it can be connected to other data, analytic framework, backup, and encryption. The **data ingest design patterns typically take the form of standard paragraphs that are used in contracts, filings (such as clinical protocols), and other legal documents**.

For example, in a memorandum of understanding, each side might specify which technical standards will be used in which parts of the project, or one side might ask the other to assert a level of cybersecurity that meets federal standards.

Depending on the data type, regulations may require informed consent, simple consent, and compliance with state and national data protection laws. These legal documents execute a basic set of functions: gathering permission from a person to take in their data into a research study, cloud platform, business, and applications. Broadly standard language dominates consumer terms of service and privacy policy, with outright cut-and-paste common from website to website. A primary goal of most of these documents is to insulate the collector from liability and the resulting documents are complex, lawyer-oriented, difficult to understand, and non-negotiable. Thus the need for user interfaces to explain data collection is increasingly accepted, with user experience designs demonstrating impact in informed consent and privacy policies.[35]

Table 3: Data collection mechanisms

Design Pattern	Description	Example
Clinical protocol	A written plan for how a health condition, drug, or device is to be studied and the procedure to be followed by the study, including technology. Connected to international and national laws in bioethics.	CDISC Clinical Trial Protocol Representation Model,[36] NIH Protocol Templates for Clinical Trials[37]
Informed consent	Standard language and user interfaces that have been vetted through legal and ethical review for specific types of data, e.g. health records or DNA	Sage Bionetworks Elements of Informed Consent,[38] GA4GH Model Consent Clauses[24]
Privacy policy	Standard language and user interfaces that have been vetted through legal and ethical review for apps and websites that collect user data	Sage Bionetworks Privacy Toolkit[39]

Design Pattern	Description	Example
Terms of service	Standard language and user interfaces that have been vetted through legal review for apps and websites that collect user data	Various auto-contract systems e.g. Formswift,[40] LegalZoom[41]

User qualification mechanisms

Data analysis governance design patterns center around how a given project admits users. While closed projects don't allow user admission at all (an anti-pattern), a broad range of scientific and consumer data holders use access committee review to decide when and if to allow a new usage of data. Committees traditionally perform both positive review ("does this researcher meet the minimum requirements?") and normative review ("is this use worthy?").[42] More recently, major biomedical collaborations began testing how broad usage rights granted objectively to users might work, especially in cloud computing.[43]

Table 4: User qualification mechanisms

Design Pattern	Description	Example
Data Access Committee (DAC)	A committee, whether a formal or informal group of individuals, with the responsibility of reviewing and assessing data access requests. Many individual groups, consortiums, institutional and independent DACs have been established but there is currently no widely accepted framework for their organization and function	dbGAP[44]
Ethics and Regulatory Review Boards	A formal accredited committee with the responsibility of reviewing and assessing research involving human beings.[45] Research institutions and governments typically run their own review boards, while for profit and other independent review boards provide the same services for a fee to the public. These boards review clinical protocols, informed consent forms, and other study related material.	AllofUs Research Program Institutional Review Board,[46] WCG-WIRB[47]
Qualified Researcher	A process by which researchers can apply to use data with a range of conditions and freedoms. Reconstructs some functionality of open licenses while allowing data download.[48]	mPower Qualified Researcher Release[49]
Data Passport	A process by which researchers can apply for general permission to do exploratory analysis of data with wide freedoms. Reconstructs some functionality of open licenses, which typically prohibits data download[50].	AllofUs Research Program Data Access Framework[43]
Registration	Data are available to users who complete a registration process and agree to terms of service. Can be used to construct a non commercial offering.	Catalog of Somatic Mutations in Cancer (COSMIC)[51]

Design Pattern	Description	Example
Open Download	Data are available on the public internet with no restrictions beyond those required to gain access to the internet and norms of attribution.	SyntheticMass FHIR Dataset,[52] Center for Medicare Services dataset downloads[53]

Mathematical governance mechanisms

A variety of technical approaches can constitute governance design patterns. De-identification is perhaps the oldest mathematical design pattern, and is incorporated into laws such as HIPAA as a key pattern of data governance. **New mathematical approaches hold promise to enable extremely broad availability and freedom at low risk**, but can be computationally expensive on enormous data sets in comparison to other forms of governance.

Table 5: Mathematical governance mechanisms

Design Pattern	Description	Example
De-identification	Language in which two or more parties agree to use either removal of fields or mathematical techniques to achieve a standard of de-identification. Typically inserted into larger governance contracts.	De-identification sample clauses[54] , HHS Sample Business Associate Agreements[55]
Blockchain	Language by which two or more parties agree to use an open, distributed, permanent ledger to broker transactions and data use.	Accord project template studio,[56] European Commission Feasability Study[57]
Homomorphic encryption	Language in which two or more parties agree to use encryption that allows computation on distributed storage and computation while preserving privacy. Most use so far in emerging smart contracts regimes.[58]	N/A
Differential privacy	Language by which two or more parties agree to publicly expose information about a dataset by describing patterns in the dataset without disclosing data about individuals. Most implementations cluster in technical projects,[59] with little standard contract language available.	N/A

Data Type

Data type — the kind of data in use — defines much of the universe of a collaboration structure. Regulation, legislation, civil rights, norms, and contracts are tied to specific kinds of data, such as electronic health records and DNA, which are in turn tied to specific kinds of risks like privacy attacks.[60] Data types are most often negotiated among collaborators, and are easiest to see inside technical standards, file formats, transactions, and more. They are not themselves governance design patterns, although design patterns will make many references to them.

In our work at Sage Bionetworks, we most often deal with data about DNA and related biological processes, "real world" data we collect about people's health through their phones and wearables, and electronic health records. There is also a vast universe of consumer data available under private contract, such as credit card or grocery data.

Pre-Existing Governance Design Patterns

We know that standard contract language can have an impact in areas outside of data governance. The Federal Demonstration Partnership (FDP), a 35-year-old cooperative initiative among 10 federal agencies and 154 institutional recipients of federal funds, implemented a vast set of standard governance documents to help groups form, apply for, and share federal funding, and report back on the collaboration outcomes.[61] An early FDP survey found that of the time that faculty committed to federal research, 42 percent was devoted to administrative activities rather than research. Researchers indicated that "grant progress-report submissions, personnel hiring, project-revenue management, equipment and supply purchases, IRB protocols and training, training personnel and students, and personnel evaluations" were the top drains on research time.[62]

The FDP years ago created what we — in this paper — call governance design patterns. Negotiations using the FDP contracts can be orders of magnitude simpler and faster because **the standard design patterns can create trust and equality among negotiators**, rather than lawyer-to-lawyer negotiation. However, a 2018 survey found a slight increase in time spent on administrative activity despite the FDP's existence.[63] This increase may be connected to the lack of support for researchers at institutions implementing negotiations.[64] Thus design patterns themselves will not solve the problem: **governance needs implementation, hiring, funding, and support to succeed**.

The FDP is joined by the SMART IRB platform funded by NIH, which eases challenges associated with initiating multisite research and provides a roadmap for institutions to implement the NIH Single IRB Review policy. Clinical protocols and consent forms often reuse language from previous projects, replicating the problem of defending past approaches whether or not they meet the current issue. This creates similar value to standardized neutral language intended for customization, i.e., design patterns. Like the FDP, SMART IRB demonstrates the value of legal tools at the design pattern level in contracts and negotiation. This finding is replicated in the design of informed consent for national research studies.

We know from the FDP and SMART IRB examples that the standardization of language and process can ease the burden of drafting contracts for co-writing grants or for regulatory paperwork. But **there is no similar unifying effort for governance design patterns (including, but not limited to legal language) in data-sharing governance.** Nor is there sufficient recognition of the complexity of implementing these tools in practice. As the 2018 FDP survey found, simplifying the language does not automatically ease the administrative burden.

Transitioning Data Across Governance Structures

Because we believe that open science makes for better science, **we work to promote governance structures and design patterns that support data sharing in a manner that is as open as possible**. But as we note above, not all data can or should be shared openly at all times.

We see value in exploring how data generated within one governance structure can migrate from a more closed context to a more open one. This allows us to define a more closed data use context for a small group of users, and export information and data to a more open structure as a form of publication. This balance is often impossible to achieve in a single governance structure: patterns that lead to reuse and patterns that lead to protection create internal conflict. The solution is to develop a program that uses a different design pattern for management of digital assets developed across the project lifecycle and/or to design a process that migrates certain data from one governance structure to the other under the right conditions. *These transitions can create balance between very real needs to protect data during analysis and the very real opportunities that can come from increased access to data*.

Our first example provided a different governance structure for sharing of data vs. sharing of analytical outputs. In this case, a private, pre-competitive group of cancer researchers working together on genetic

subtypes for colorectal cancer perceived that sharing data with one another was worth the risk, but publishing the data — or the algorithms they build with that data — onto the public internet, under an open license, was not.[65] This project used a club structure to enable data sharing so that each researcher could develop their algorithm with the extended set of data but, in exchange, the group retained the right to compare those algorithms and develop a consensus model that could be published openly. This shift in design patterns supports a transition from club to open structure for key digital assets in a manner that balances the public benefit of the collective wisdom with the private incentives of the individual researchers.

Figure 8: Transitioning from one governance structure to another requires transitioning from one governance design pattern to another.

A second example illustrates a shift from a trusted environment to an open structure for data sharing over time. The Accelerating Medicines Partnership for Alzheimer's Disease, a public private partnership between the US NIH and the pharmaceutical industry, engaged academic and pharmaceutical industry researchers in a pre-competitive program to identify new drug targets in Alzheimer's from multi-omic brain data. The researchers agreed to share data, algorithms and outcomes with each other immediately upon data generation — and to periodically migrate those assets into the public domain upon a pre-specified schedule (ADknowledgeportal.org).

In a third example, the Breast Cancer Surveillance Consortium and the Icahn School of Medicine at Mount Sinai had a database of 640,000 mammograms. Wondering if machine learning might provide assistance to, or even match, expert radiologists, they hoped to convene a computational contest. But the data had not been consented for either high freedom or high availability, and bore risks of re-identification of the patients represented by the data. Beyond that, the scale of the data made broad access to it computationally and economically expensive. This required an analysis structure that was tightly controlled, but some form of data distribution that was liberal enough to support hundreds of data users.[66]

By creating a synthetic form of the data set — which could be downloaded widely, as it does not represent real people — a wide set of analysts could use this data to inform algorithm development. Acting as a steward, Sage Bionetworks could then run models on behalf of the submitters and return results and feedback to allow for model tuning, iteration, and development. This transition of synthetic data from the model-to-data structure to open data kept an "air gap" around the real data to protect its privacy, but allowed hundreds of users worldwide to build algorithms.[5]

Illustrative use cases

The COVID19 pandemic has forced an unprecedented spike in scientific data collaboration.

Our approach of using governance design patterns to rapidly build desired governance structures has, so far, performed well under time pressure. We illustrate here two use cases, the National COVID Cohort Collaborative and the COVID Recovery Corps, to demonstrate how we could stand up two different projects in less than two months without sacrificing legal quality or ethical rigor.

National COVID Cohort Collaborative (N3C)

The National COVID Cohort Collaborative (N3C) is a collaboration among 60 grant-receiving clinical centers and their partners, 5 pre-existing distributed health data networks, the NIH's National Center for Advancing Translational Sciences (NCATS), and the pre-existing National Center for Data To Health project (CD2H). The N3C's goal is to enable detailed data analysis of clinical records for COVID patients for a wide variety of users while preserving individual privacy of patients.

Formed rapidly in response to the outbreak of the pandemic, the N3C is assembling relevant electronic health record data of COVID-19 patients from participating institutions. The data live in a secure analytical environment where tools and algorithms can be rapidly evaluated, and clinicians and researchers can ask granular and complex clinical questions. Additionally, the same data will drive a synthetic data set available for broad download and reuse, and partners will expose their data for pre-written federated queries.

As members of the CD2H project, Sage Bionetworks governance leaders chose a suite of governance structures and transition: first, a club managed by NCATS into which data flows from the many data owners. Second, the N3C creates a range of transitions for data. Some users sign a data use agreement to get into the "real" data set, others to run federated queries. Synthetic data will be available for registered download. The entire process is run through a single ethical review board at Johns Hopkins University under SMART IRB.

By using governance structures and design patterns, the N3C was able to move governance rapidly while retaining high quality. Within two months of launch, the N3C protocol passed review at Johns Hopkins, more than 37 institutions have signed the single data transfer agreement, and data access is already in testing. Our governance aims to enable the project to achieve its goals to improve the efficiency and accessibility of analyses with COVID-19 clinical data, expand our ability to analyze and understand COVID, and demonstrate a novel approach for collaborative pandemic data sharing.
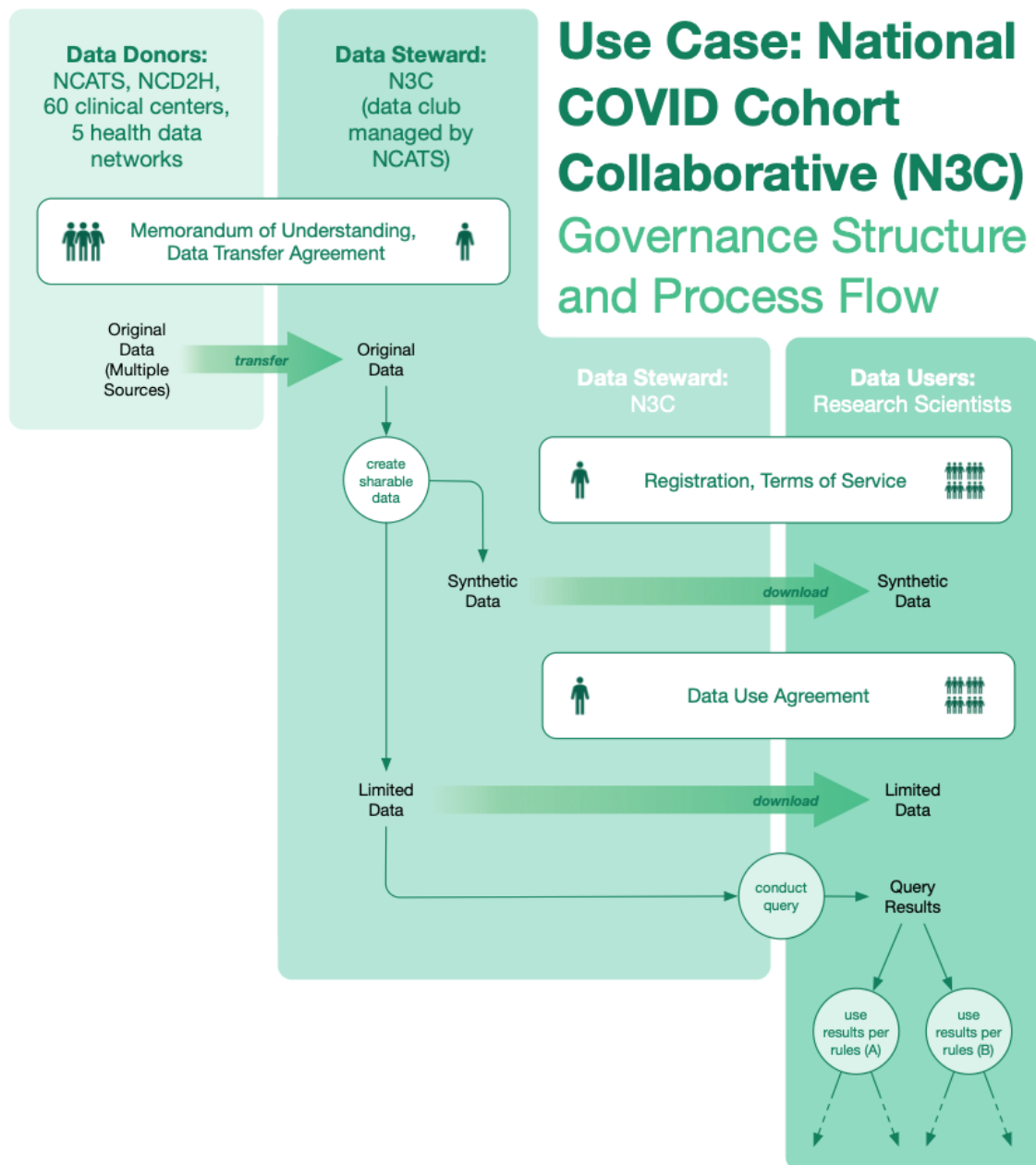
Figure 9: How N3C Data Governance is Structured.

COVID Recovery Corps (CRC)

The Covid Recovery Corps (CRC) is a collaborative research study by Columbia University and Sage Bionetworks, supported by the Chan Zuckerberg Initiative. The CRC aims to enroll people from the New York City metro area and have recovered from COVID-19 to partner with scientists to better understand the recovery process and long term effects from the disease.

Some patterns of the CRC are stable governance design patterns: it's a research study feeding data into a club under clinical protocols and using informed consent. The study also launches with a 50,000 person Facebook community, anticipates national at-home testing, and its protocol must contemplate cities other than New York joining the protocol as new "nodes" on a clinical network.

Using governance structures and design patterns, less than 50 days elapsed between kickoff meetings and IRB approval for the project.

Figure 10: How CRC Data Governance is Structured.

Governance and Design In Context

Our approach in this Green Paper sits within a broader movement to apply human-centered design to policy. The concepts of structures and patterns will be familiar to designers. Part of our argument is that governance needs design, and that designers should embrace governance as a topic area.

A well-designed governance system must center empathy for those using it, and frame the design work in terms of their experience of use. While governance creation often starts with desired outcome (e.g., "more fairness" or "more ethics" or "faster outcomes" or "collaboration formed") it is rare that those outcomes include or articulate aspects of the lived experience of technology-based biology/data science. Designing governance systems for data science, like other digital services, should begin by exploring and pinpointing the needs of the people who will use the service, and the ways the service will fit into their lives.

One helpful tool for human-centered design is the persona. Based on interviews with a wide variety of users to understand their views and experiences, it is possible to derive a handful of patterns in the types of users, each augmented with vivid details — a name, picture, personal history, etc — to conjure an image of the person being designed for. Personas provide user archetypes that capture the essentials and are easy to recall and reference throughout the design process. In the context of data use, personas could allow us to test a governance feature with respect to the needs of a variety of stakeholders in open science.

Scenarios represent another tool. Like a persona, a scenario is an artifice that is based on empirical research, analysis of patterns, and synthesis of attributes, embellished with colorful, memorable details to aid in functionality. However, instead of creating a set of hypothetical people, we are creating a set of hypothetical contexts in which they would act. Like personas, scenarios define the space of the possible, and are useful only as a set (not individually) to explore the robustness of particular options. Also, a hallmark of human-centered design is frequent, rapid, iterative prototyping, which incorporates user feedback back into the design process

as often as possible. Testbeds to enable frequent, rapid, iterative prototyping and integrate user feedback are another often overlooked, but entirely necessary, component of the governance enterprise.

Conclusion

Data use collaborations are incredibly diverse in terms of the types of artifacts being shared, the degree to which they are shared, the purposes with which they are shared, not to mention the power dynamics among collaborators. This means that we must attend to the design of the rules and incentive structures that govern human behavior as much as we attend to the design of hardware and software. While there is no one-size-fits-all solution, we can learn from observed research collaborations and reuse patterns of governance systems that worked well for those collaborations when we design new systems of governance, or transition from one system to another. The important thing is that form follows function: governance must be determined by the nature of collaboration desired. For example, large projects require governance campaigns and design, are high impact and rare, and often possess long feedback loops before value creation. Small projects require on-the-ground governance and immediate feedback loops.

There are limits to the structure-design pattern approach. One is that implementation is messy in reality: there will inevitably be some trial and error, as technology and governance are co-iterated in practice. Exactly how technology and governance combine in actual use will be determined when the project runs. Additionally, institutions within science have longstanding incentives to perform customized governance as a service, which can make traction for new methods difficult. Funder mandates have been quite successful in open access governance and may be necessary in data governance as well.

Thus by our model we do not mean there is one "right" governance design, but instead that successful designs must share two traits: **they should allow collaborators to get going quickly in a legally valid structure, and they should also be flexible enough to change to meet the needs of collaborators** — especially in the early stages of design or transition. Turning ad hoc processes into governance design patterns means they can be encoded in software: services and products can integrate some key patterns and variables, supporting implementation in daily practice.

Our structure-design pattern for data governance pairing sits inside a growing understanding of how the ways that human-centered design can help in policymaking and governance. In addition to this Green Paper, which will map out the landscape of data governance and its mixture into technical platforms, we are in early development for three major projects that take advantage of that mixture. Each draws on Sage's goals of increasing the scope of responsible sharing, of representative data, to increase the reliability of claims that come out of data analysis.

Like our structure-design pattern work here, the next stage builds on the way we develop software and other products using human-centered design. The work of using open science to generate reliable claims from representative data (and of sharing both responsibly) remains laborious, manual, and time consuming. We propose to extend data governance into software testing frameworks and information architecture to address this problem.

Reliable Analytic Test Framework

Our Reliable Analytic Environment project looks to extend the concept of automated testing frameworks from software testing to create similar functionality to evaluate the reliability of analytical outputs. Instead of looking for software bugs, we will look for reliability bugs and their impacts. These could come from small sample sizes, failure to implement known best practices, and failure to test against known benchmarks.

Representative Algorithm Test Framework

We propose to create a testing framework for representativeness in medical research. Drawing on research into social determinants of health and the rapidly increasing public health data sources available, we can begin

to craft software services that take a person's algorithmic designs and test them against public, semi-public, and potentially even semi-private data sets to evaluate the likelihood of under-sampling bias. This will allow data scientists to examine how their code looks beyond their non-representative samples, and see if they are accidentally encoding bias in their work — and, if so, what the impacts are.

Information Architecture of Data Governance

For areas such as informed consent, privacy policy, contract negotiation, and data licensing, simply generating and releasing tools and licenses has not been shown to have a nonlinear impact. One reason is that an enormous amount of tacit knowledge is required to use these kinds of tools appropriately, and they are often used in spaces that are distant from the data scientist (i.e., lawyers discuss these tools, not coders). Our IA work will explore how we can learn from libraries and other complex information spaces, so that we can find areas where we need to build tools. Tools could include pedagogical resources (up to and including online courses), software services, and testing environments.

Epilogue

We envision the models of governance we describe here as part of a fundamental infrastructure for a robust science commons. This marries the development of technologies for open science with the development of technologies for their governance. In our vision, the science commons infrastructure would connect scientific knowledge — at least in part — across domains and guide the evolution of data science across scholarly paradigms. Every scientist would be afforded access to digital services with which to contribute to and benefit from a scientific knowledge commons and be able, in turn, to manage those services legally and ethically. For assets from data to algorithms to protocols to prose to diagrams, anyone could store, transfer, compute, collaborate, publish, read, verify, and critique. This infrastructure will require the hardware and software to create virtual spaces for data sharing, as well as the governance systems to make data sharing effective, efficient, and responsible.

Effective governance mechanisms would incentivize high quality scientific work and protect against abuses. For example, virtual spaces could be certified for different purposes: zones of complete openness, zones of privacy, zones that are mixed, and experimental zones that allow for work beyond default rules. Across spaces with different rules, it may also make sense to allow for selective enforcement to stimulate innovative activity in virtual "bohemias." Such spaces could benefit emergent, quasi-scientific domains (e.g. DIY biology), which need more structure and connection to established fields than they presently have, but not so much that it dulls the novel nature of the work. Like hardware and software, governance would have to be tailored and evolve over time.

Creating a digital infrastructure for the science commons is no small undertaking. But science has always faced a tragedy of the commons. Vannevar Bush wrote in his now infamous text, *Science: the Endless Frontier*: "there are areas of science in which the public interest is acute but which are likely to be cultivated inadequately if left without more support than will come from private sources." For the last 75 years, the federal government has stepped in to address this market failure through the National Science Foundation, a constellation of National Laboratories, and other means. "Basic research leads to new knowledge. It provides scientific capital. It creates the fund from which the practical applications of knowledge must be drawn."[67]

In our science-innovation mythos, the commons was understood narrowly as fundamental scientific knowledge — the "seedcorn" of innovation — and the tragedy of its underproduction was to be solved by public funding of discipline-based, investigator-driven, basic research. But the world has changed over the last 75 years. Notably, there is no longer a market failure for R&D. Private funding of R&D has surpassed federal funding since 1980 and that gap has been steadily growing,[68] and "innovations often occur that do not require basic or applied research or development."[69] Despite the growth of scientific knowledge over the past decades, economic productivity has stagnated; more R&D at universities does not automatically translate into

more value in the economy.[70] Something is missing. The scientific commons has not disappeared, but it has shifted … and so has the tragedy.

The old commons tragedy was the production of *knowledge*, with the focus on the product: our metaphorical seed corn. The new commons tragedy is the *production* of knowledge, with the focus on the production: our metaphorical system of agriculture. Seed corn cannot grow if the topsoil has been eroded, or if a drought has reduced the amount of available water, or if a shifting climate has increased the number or freezing degree days, or if bollweevils eat all of our seedcorn because, though monoculture, our stock was not diverse enough. If collecting more seedcorn is our only tactic, then we may go hungry when our environmental context changes. Instead, we must have a more diverse set of tactics within an overarching, robust strategy of cultivation.

How will the practice of science, itself, operate for the good of society in our new era? We must reconceive the scientific commons as a system and, therefore, pay more attention to what lies between individual sciences: the transdisciplinary infrastructure and boundary-spanning institutions that enable the entire system of knowledge production to function. We must build the tools and practices for integrating different communities of knowledge workers into a larger ecology of knowledge in order to make the system as a whole more productive and more resilient.

Good governance for everyone is where we propose to begin.

Acknowledgments

References

1. Vaidhyanathan S. Antisocial media: how facebook disconnects US and undermines democracy. New York, NY, United States of America: Oxford University Press; 2018.

2. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency. 2018 Jul 16. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

3. Knorr-Cetina K. Epistemic cultures: how the sciences make knowledge. Cambridge, Mass: Harvard University Press; 1999.

4. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, Lotter W, Jie Z, Du H, Wang S, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Network Open. 2020;3(3):e200265. https://doi.org/gg2dvg. doi:10.1001/jamanetworkopen.2020.0265

5. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nature Biotechnology. 2018;36(5):391–392. https://doi.org/gg2dvh. doi:10.1038/nbt.4128

6. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey ER, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific Data. 2016;3(1). https://doi.org/gg2dvj. doi:10.1038/sdata.2016.11

7. Garrison NA. Genomic Justice for Native Americans. Science, Technology, & Human Values. 2012;38(2):201–223. https://doi.org/gg2dvk. doi:10.1177/0162243912470009

8. Cheah PY, Tangseefa D, Somsaman A, Chunsuttiwat T, Nosten F, Day NPJ, Bull S, Parker M. Perceived Benefits, Harms, and Views About How to Share Data Responsibly. Journal of Empirical Research on Human Research Ethics. 2015;10(3):278–289. https://doi.org/f7pmw8. doi:10.1177/1556264615592388

9. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1). https://doi.org/bdd4. doi:10.1038/sdata.2016.18

10. Pasquetto IV, Randles BM, Borgman CL. On the Reuse of Scientific Data. Data Science Journal. 2017;16. https://doi.org/gf87pf. doi:10.5334/dsj-2017-008

11. Bardill J, Garrison NA. Naming Indigenous Concerns, Framing Considerations for Stored Biospecimens. The American Journal of Bioethics. 2015;15(9):73–75. https://doi.org/gg2jdp. doi:10.1080/15265161.2015.1062164

12. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society. 2018 Feb. https://datasociety.net/wp-content/uploads/2018/02/DataSociety_Fairness_In_Precision_Medicine_Feb2018.pdf

13. Ross MW, Iguchi MY, Panicker S. Ethical aspects of data sharing and research participant protections. American Psychologist. 2018;73(2):138–145. https://doi.org/gc4823. doi:10.1037/amp0000240

14. McNealy J, Mullis MD. Tea and turbulence: Communication privacy management theory and online celebrity gossip forums. Computers in Human Behavior. 2019;92:110–118. https://doi.org/gg2x6m. doi:10.1016/j.chb.2018.10.029

15. Wilkins CH, Schindler SE, Morris JC. Addressing Health Disparities Among Minority Populations. JAMA Neurology. 2020;77(9):1063. https://doi.org/gg2jdn. doi:10.1001/jamaneurol.2020.1614

16. Collins FS. The Human Genome Project: Lessons from Large-Scale Biology. Science. 2003;300(5617):286–290. https://doi.org/d2zg2r. doi:10.1126/science.1084564

17. Schwarte A, Haase P, Hose K, Schenkel R, Schmidt M. FedX: Optimization Techniques for Federated Query Processing on Linked Data. In: Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2011. https://doi.org/bgnp5d. doi:10.1007/978-3-642-25073-6_38

18. Lane J, Shipp S. Using a Remote Access Data Enclave for Data Dissemination. International Journal of Digital Curation. 2008;2(1):128–134. https://doi.org/dfkk4c. doi:10.2218/ijdc.v2i1.20

19. Suver C, Thorogood A, Doerr M, Wilbanks J, Knoppers B. Bringing Code to Data: Do Not Forget Governance. Journal of Medical Internet Research. 2020;22(7):e18087. https://doi.org/gg2x6p. doi:10.2196/18087

20. Greshake Tzovaras B, Angrist M, Arvai K, Dulaney M, Estrada-Galiñanes V, Gunderson B, Head T, Lewis D, Nov O, Shaer O, et al. Open Humans: A platform for participant-centered research and personal data exploration. GigaScience. 2019;8(6). https://doi.org/gg2jdq. doi:10.1093/gigascience/giz076

21. Ostrom E. Beyond Markets and States: Polycentric Governance of Complex Economic Systems. American Economic Review. 2010;100(3):641–672. https://doi.org/dhj38d. doi:10.1257/aer.100.3.641

22. McDonald S. Reclaiming Data Trusts. Centre for International Governance Innovation. [accessed 2020 Jun 18]. https://www.cigionline.org/articles/reclaiming-data-trusts

23. Torous J, Nebeker C. Navigating Ethics in the Digital Age: Introducing Connected and Open Research Ethics (CORE), a Tool for Researchers and Institutional Review Boards. Journal of Medical Internet Research. 2017;19(2):e38. https://doi.org/f9pmgs. doi:10.2196/jmir.6793

24. Dyke SOM, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, Brookes AJ, Spalding JD, Thompson M, Roos M, et al. Consent Codes: Upholding Standard Data Use Conditions. PLOS Genetics. 2016;12(1):e1005772. https://doi.org/gg2d32. doi:10.1371/journal.pgen.1005772

25. Share your work. Creative Commons. [accessed 2020 Oct 6]. https://creativecommons.org/share-your-work/

26. Open Data Commons Open Database License (ODbL) — Open Data Commons: legal tools for open data. [accessed 2020 Oct 6]. https://opendatacommons.org/licenses/odbl/

27. microsoft/Open-Use-of-Data-Agreement. Microsoft; 2020. https://github.com/microsoft/Open-Use-of-Data-Agreement

28. Health Care Systems Network. (n.d.). [accessed 2020 Jun 18]. http://www.hcsrn.org/en/Tools%20&%20Materials/GrantsContracting/HCSRN_DUAToolkit.pdf

29. Standard Contractual Clauses (SCC). European Commission - European Commission. [accessed 2020 Oct 6]. https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/standard-contractual-clauses-scc_en

30. Material Transfer Agreements (MTAs) for Technology | AUTM. [accessed 2020 Oct 6]. https://autm.net/surveys-and-tools/agreements/material-transfer-agreements/mta-toolkit/uniform-

biological-material-transfer-agreement

31. A data processing addendum for the CCPA? [accessed 2020 Oct 6]. https://iapp.org/news/a/a-data-processing-addendum-for-the-ccpa/

32. Sample Addendum Addressing Article 28 GDPR and Incorporating Standard Contractual Clauses for Controller to Processor Transfers of Personal Data. [accessed 2020 Oct 6]. https://iapp.org/resources/article/sample-addendum-addressing-article-28-gdpr-and-incorporating-standard-contractual-clauses-for-controller-to-processor-transfers-of-personal-data/

33. Creative Commons — CC0 1.0 Universal. [accessed 2020 Oct 6]. https://creativecommons.org/publicdomain/zero/1.0/

34. Open Data Commons Public Domain Dedication and License (PDDL) — Open Data Commons: legal tools for open data. [accessed 2020 Oct 6]. https://opendatacommons.org/licenses/pddl/

35. Moore S, Barone V, Bot B, Doerr M, on behalf of the All of Us Consent Work Group and the Low Health Literacy Consent Study research team. Informed Consent for the All of Us Research Program. Forthcoming/in press. 2020.

36. PRM | CDISC. [accessed 2020 Oct 6]. https://www.cdisc.org/standards/foundational/protocol

37. Protocol Templates for Clinical Trials | grants.nih.gov. [accessed 2020 Oct 6]. https://grants.nih.gov/policy/clinical-trials/protocol-template.htm

38. Elements of Informed Consent. Sage Bionetworks. [accessed 2020 Oct 6]. https://sagebionetworks.org/tools_resources/elements-of-informed-consent/

39. What is a Privacy Toolkit? | Design System Template. [accessed 2020 Oct 6]. https://designmanual.sagebionetworks.org/privacy_toolkit.html

40. FormSwift: Create Legal Documents. [accessed 2020 Oct 6]. https://formswift.com/

41. LegalZoom: Start a Business, Protect Your Family. LegalZoom. [accessed 2020 Jun 18]. https://www.legalzoom.com

42. Cheah PY, Piasecki J. Data Access Committees. BMC Medical Ethics. 2020;21(1). https://doi.org/gg2d3z. doi:10.1186/s12910-020-0453-z

43. Data Use Policies – All of Us Research Hub. [accessed 2020 Oct 6]. https://www.researchallofus.org/data-use-policies/

44. National Center for Biotechnology Information. (n.d.). DbGaP Request Procedures to Access Individual-Level Data. 2011 Mar 23. https://dbgap.ncbi.nlm.nih.gov/aa/dbgap_request_process.pdf

45. Mazur DJ. Evaluating the science and ethics of research on humans: a guide for IRB members. Baltimore: Johns Hopkins University Press; 2007.

46. All of Us Research Program IRB | National Institutes of Health. [accessed 2020 Oct 6]. https://allofus.nih.gov/about/who-we-are/institutional-review-board-irb-of-all-of-us-research-program

47. WIRB Home. [accessed 2020 Oct 6]. http://www.wirb.com/Pages/Default.aspx

48. Grayson S, Suver C, Wilbanks J, Doerr M. Open Data Sharing in the 21st Century: Sage Bionetworks' Qualified Research Program and Its Application in mHealth Data Release. SSRN Electronic Journal. 2019. https://doi.org/gg2jds. doi:10.2139/ssrn.3502410

49. Wilbanks J, Friend SH. First, design for data sharing. Nature Biotechnology. 2016;34(4):377–379. https://doi.org/gbrb37. doi:10.1038/nbt.3516

50. All of Us Research Program Begins Beta Testing of Data Platform | National Institutes of Health (NIH) — All of Us. [accessed 2020 Oct 6]. https://allofus.nih.gov/news-events-and-media/announcements/all-us-research-program-begins-beta-testing-data-platform

51. Wellcome Sanger Institute: Catalogue of Somatic Mutations in Cancer Data Downloads. [accessed 2020 Jun 18]. https://cancer.sanger.ac.uk/cosmic/download?genome=37

52. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association. 2018;25(3):230–238. https://doi.org/gb4mkg. doi:10.1093/jamia/ocx079

53. Dataset Downloads | CMS. [accessed 2020 Oct 6]. https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads

54. De-identified Information Sample Clauses. [accessed 2020 Oct 6]. https://www.lawinsider.com/clause/de-identified-information

55. Office for Civil Rights (OCR): Business Associate Contracts. HHS.gov. 2008 May 21 [accessed 2020 Jun 18]. https://www.hhs.gov/hipaa/for-professionals/covered-entities/sample-business-associate-agreement-provisions/index.html

56. Template Studio. Accord Project. 2019 Jun 3 [accessed 2020 Oct 6]. https://accordproject.org/projects/template-studio/

57. Modelling the EU Economy as an Ecosystem of Contracts. Volume 1, feasibility study. 2020. doi:10.2838/28184

58. Nugent T, Upton D, Cimpoesu M. Improving data transparency in clinical trials using blockchain smart contracts. F1000Research. 2016;5:2541. https://doi.org/ggzsn4. doi:10.12688/f1000research.9756.1

59. OpenDP. GitHub. [accessed 2020 Oct 6]. https://github.com/opendifferentialprivacy

60. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nature Genetics. 2020;52(7):646–654. https://doi.org/gg4dcj. doi:10.1038/s41588-020-0651-0

61. Home - The Federal Demonstration Partnership. [accessed 2020 Oct 6]. https://thefdp.org/default/

62. Decker R, Wimsatt L, Trice A, Konstan J. A profile of federal-grant administrative burden among Federal Demonstration Partnership faculty. 2007. https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_054586.pdf

63. Schneider SL. Results of the 2018 FDP Faculty Workload Survey: Input for optimizing time on active research. 2019. http://thefdp.org/default/assets/File/Presentations/Schneider%20FDP%20FWS3%20Results%20Plenary%20Jan19.pdf

64. Michel KK. IDENTIFICATION OF PRIMARY TIME DRIVERS IN COMPLIANCE FOR NON-MEDICAL STEM RESEARCH, AND BEST PRACTICES FOR REDUCING TIME INVESTMENT BY FACULTY [Thesis]. Johns Hopkins University; 2019. http://dspace-prod.mse.jhu.edu:8080/handle/1774.2/62063

65. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine. 2015;21(11):1350–1356. https://doi.org/f7xmph. doi:10.1038/nm.3967

66. Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, Fuchs C, Gofer E, Kumar L, Wolfinger RD, et al. A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. JCO Clinical Cancer Informatics. 2017;(1):1–15. https://doi.org/gg2jdr. doi:10.1200/cci.17.00018

67. Bush V. Science the Endless Frontier. 1945 [accessed 2020 Jun 18]. https://www.nsf.gov/od/lpa/nsf50/vbush1945.htm

68. Figure 4-1 - S&E Indicators 2018 | NSF - National Science Foundation. [accessed 2020 Oct 6]. https://www.nsf.gov/statistics/2018/nsb20181/figures/fig04-01

69. Congressional Research Service. Industrial Competitiveness and Technological Advancement: Debate Over Government Policy. 2012. https://www.everycrsreport.com/files/20121203_RL33528_ebc4cd402342b9bbf91bebdb37bbe6bdc06e57ae.pdf

70. Arora A, Belenzon S, Patacconi A, Suh J. The Changing Structure of American Innovation: Some Cautionary Remarks for Economic Growth. National Bureau of Economic Research. 2019 Jun 28. https://doi.org/gg2jdt. doi:10.3386/w25893