Cell
Leading Edge

Commentary

A “data sharing trust” model for rapid, collaborative science

Vincent Chan,[1,2,3] Pier Federico Gherardini,[6] Matthew F. Krummel,[2,3,*] and Gabriela K. Fragiadakis[3,4,5,*]
[1]Department of Microbiology and Immunology, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143-0511, USA
[2]Department of Pathology, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143-0511, USA
[3]ImmunoX Initiative, Department of Microbiology and Immunology, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143-0511, USA
[4]UCSF CoLabs, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143-0511, USA
[5]Division of Rheumatology, Department of Medicine, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143-0511, USA
[6]Parker Institute for Cancer Immunotherapy, 1 Letterman Drive, Suite D3500, San Francisco, CA 94129-1504
*Correspondence: matthew.krummel@ucsf.edu (M.F.K.), gabriela.fragiadakis@ucsf.edu (G.K.F.)

https://doi.org/10.1016/j.cell.2021.01.006

SUMMARY

Complex datasets provide opportunities for discoveries beyond their initial scope. Effective and rapid data sharing and management practices are crucial to realize this potential; however, they are harder to implement than post-publication access. Here, we introduce the concept of a “data sharing trust” to maximize the value of large datasets.

Collaborative science in the era of complex multi-modal datasets

With the advent of new technologies and an appreciation for systems-level analyses, there are a growing number of research endeavors that generate large, multi-modal datasets. These projects often involve many investigators who bring complementary expertise in biological sub-specialties, both in generating and analyzing specific data types, and in contributing their clinical perspective and understanding. Such projects present an incredible opportunity for scientific advancement, but to be successful, they require rapid iteration, elaboration, and sharing in near real-time, often beyond the planned duration and scope of the initial project. A key development that the current coronavirus disease 2019 (COVID-19) pandemic has brought to the forefront is the importance of near real-time data sharing—bringing many eyes and many insights to important questions. Furthermore, the NIH recently released a ‘‘Data Sharing and Management Policy’’ requiring a stated data sharing and management plan for all federally funded projects; this both underscores the importance of this practice and prompts the research community to devise practical solutions to this challenge.

Here, we present a perspective on possible approaches that move beyond the traditional ‘‘access-restriction’’ models, which are often limited to data sharing with an emphasis on secondary analysis, typically after a first publication has already been generated. These models are inflexible, and they tend to overvalue the work involved in the production of raw data and undervalue analytical work and interpretation. Instead, we will present a “data sharing trust” model that seeks to honor the personal incentives that drive the passion of scientists while enabling the community to access well-annotated data as early as, and ideally in concert with, its production. We will highlight our application of these ideas in our COVID-19 research effort, a collaboration and data sharing trust across over 150 researchers. These are ideas that need further development but might represent the right seed to make data sharing a valuable enterprise for both investigators and institutions, even beyond our current state of pandemic-driven, community-minded projects.

Barriers to data sharing

Big and fast projects emphasize a need for data sharing that is concurrent with its production, quality-control, and primary insight generation. However, there are currently several barriers to fluid and timely data sharing among researchers. Beyond logistical constraints, including a lack of infrastructure for efficient data capture and sharing and the significant time and effort required for researchers to curate the datasets, there are three significant stakeholders to consider when crafting agreements for collaboration and data sharing:

(1) Investigators. Publications are necessary for career advancement, and investigators seek to make contributions that solidify their status in their field; data sharing could jeopardize one’s chances for publication or to be credited with an important discovery.
(2) Research Institutions. Institutions own and monetize the intellectual property (IP) developed by their investigators. Data sharing can lead to other parties developing IP based on data produced by their

investigators, therefore weakening the IP position of the institution.

(3) Human Subjects. For research involving human subjects, data sharing is strictly regulated to protect patient privacy. Failure to properly address these concerns undermines public trust, jeopardizes participation in future research, and could result in serious and costly legal consequences.

The latter two barriers are often derived from the investigator's home institution and will not be extensively covered here. Briefly to the third, investigators must take special care in drafting informed consent forms that enable, as much as possible, usage of the data for future research purposes—including a request for consent to distribute de-identified data to additional investigators, potential industry partners, and/or public repositories—with secure systems in place to shield protected health information and to respect patient privacy and autonomy. Institutional Review Boards should be engaged early in the design of research programs to ensure that protocols can be developed to maximize data sharing while at the same time protecting patient interests.

The main focus of our perspective is the first barrier: the realities of a career in academic science that expects publications—most importantly, first and corresponding authorships—of novel findings in high-impact journals. This reality is complemented by the very real focus of scientists on having a reasonable window of time to discover in the data what it is they sought out to study in the first place. It is therefore an oversimplification to expect or mandate immediate broad release of an investigator's data without addressing these realities in some way. In addition to data generation, the primary investigators invest substantial time and resources into conceiving and organizing a study. Their main motivation, producing scientific discoveries and breakthroughs, requires time for them to analyze and expand upon the data to publish derived insights. For investigators to want to share their data, there must be trust that others, using the fruits of their labor, will adequately include them in both the discovery process and in the assignment of credit for their works. Incentives might be formulated to convince scientists to share against their perceived interests (Bierer et al., 2017; Olfson et al., 2017), but these do not always provide an environment that minimizes the risk of being "scooped from within" using your own data.

Limitations of current data sharing models

Two contrasting models dominate current data sharing practices in the biological sciences: (1) data sharing and distribution at the time of publication, and (2) real-time or near-real-time data release to the public (Birney et al., 2009). Both of these models should be seen as important progress in norms and practices toward data sharing and collaborative science. Beyond the NIH mandate, other funding agencies and journals are requiring a commitment to data sharing (Sim et al., 2020), and a wealth of data thus exists in the public domain for broad investigation and use.

However, these models present their own limitations that hinder investigators. In the first, tying data sharing requirements to publication incurs a large time-delay between data generation and sharing, ranging from months to years. During that time, other investigators could have been deriving additional insight from the data. Furthermore, this likely would have been the ideal time for fruitful collaboration because this is the time window when the primary investigator is most intently focused on this particular dataset. Preprint servers like BioRxiv, MedRxiv, and ChemRxiv might accelerate this timeline, but rarely include full data release. In addition, requirements and systems for data sharing are often very burdensome for the researcher, resulting in labs keeping "two sets of books:" an internal version of the data that they rely on for analysis and an external version with minimal curation that they are required to release after-the-fact. This second set might not include the raw data and might be less granular and less thoroughly annotated than their "in-house" version. This poses a large problem for new researchers who would like to work with this data and, though surmountable, results in more work and lower data integrity. A system that curates the data into a well-annotated "data library" at the time of collection is therefore optimal.

Largely in response to these shortcomings, others advocate for the opposite approach: near-real-time data release. Although this approach solves the time-delay problem, it is problematic for both investigators and institutions, as discussed in the previous section. Ultimately, a more nuanced model is necessary to ensure timely dissemination of data and encourage collaboration while at the same time safeguarding the interests of all the stakeholders in the research ecosystem.

A data trust for real-time data sharing

We present here a perspective on possible approaches that bridge the divide between these two camps of thought and move beyond traditional access-restriction models. We begin with the concept of "shells" for data sharing (Figure 1A), starting with the data producers (or data "stewards"), who generate and share the data freely across collaborators (the other stewards and their labs), followed by restricted and subsequent data access to the broader participating institutions, and finally to the public. We also present the idea of a succession from "raw" to "processed" to "insight-level" data categories, with the first two categories being shared in near real-time, whereas the last category is dependent on initial insight-generating studies and/or analyses (Figure 1B). We note that in each transition from raw to curated insight-level data, investigators continually add value by contributing to the data curation and interpretation process, distilling ideas into insights in a steady progression. There are often many levels to which data can be processed to generate insights. Similar to current embargo models, each data class would then also have a proposed data release schedule determined by the steward and project leadership.

Our underlying ethos for data sharing is that it is essential to define a human trust for all investigators who see data within a sphere ("data sharing trust"). For example, all collaborating investigators can see the data generated as part of the project, but each investigator must first contact the stewards of that particular dataset and engage them as collaborators before using the data. Such engagement should include routine reporting of insights, and investigators who subsequently use data in any publication are obliged to, at minimum, offer authorship to the stewards (primary

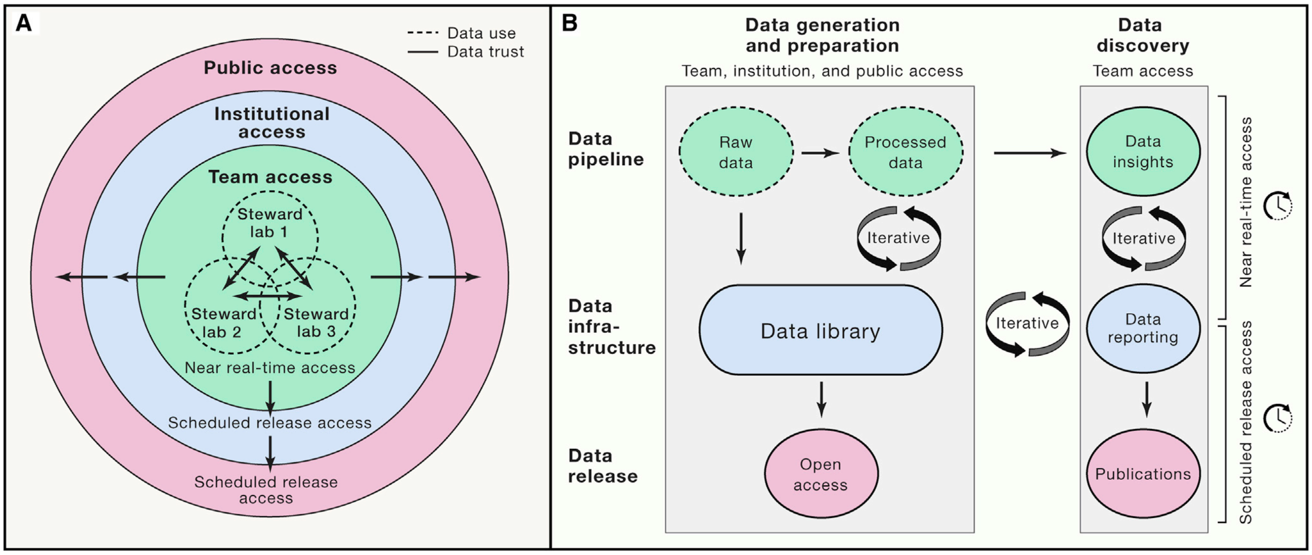

Figure 1. A model for real-time data sharing based on data trust
(A) Shells of data sharing. The inner-most circle represents collaborative team sharing—all collaborators have access to the data but recognize the importance of the lab stewards in being contacted and continually included in data use by the rest of the team. At each level, codes of conduct for data sharing can be defined and this will provide researchers confidence to do their work within the sphere rather than taking data, and insight, outside the data sharing model. Notably, a team can be within or across institutions though all must adhere to the agreed upon terms of the data trust.
(B) Logistics of data sharing during insight generation. Integral to an agreement on data sharing is recognizing the various types of data that are generated. Expectations around upload and access for each data type should be considered. Here, we distinguish between (1) "raw data," for example raw .fastq files from biological sequencing; (2) "processed data," the initial curation of the data into a useable form, such as the generation of gene counts matrices from sequencing files; and (3) "data insights," the biological insights and understanding derived from the data. Data in the data sharing trust is shared in the team's data library, and insights are reported to team members. At a later date, data are released to those outside of the team via open access tools and publications.

investigators) and all contributing members of that dataset. New collaborators interested in accessing data must then also agree to follow the proposal submission process and read and sign a trust agreement prior to data access. Notably, although a team can be either intra- or inter-institutional, the model requires an established and ongoing trust, which could limit the total number of collaborators. Despite this, the codification and agreement of trust allows access to be granted to many more investigators than would typically be possible if a trust agreement was not settled in advance. At each level, the nature of trust changes, and it is important for all parties to understand and agree to the costs and benefits of including more access.

Data sharing among collaborating researchers

We propose a system of data sharing that promotes both data integrity and collaborative discovery, as summarized in Figure 1B. In our scheme, raw and processed data are deposited in a shared data trust platform, or "data library," in near real-time and ideally directly from the instruments. By setting the expectation that the dataset is immediately worthy of curation into a shareable format, our framework guarantees that QC and data integrity are conceived as important from the start. Sharing data early on has the additional benefit of informing project researchers that the data are available for collaboration and offers the possibility for integrative analysis with the consent of the stewards.

Large enough incentives need to exist to strongly motivate or encourage participating investigators to go through the effort of curating and depositing data into the system. Several examples of data platform features that can incentivize its use and create added value for individual investigators include, but are not limited to:

(1) Support from a data science team that manages the platform, including standardization and quality-control of the data as part of the import process;
(2) Development of data visualization and analysis tools on top of the platform, that can accelerate the discovery process once data have been imported, rather than requiring re-exporting to yet another platform where insights will be private;
(3) Seamless integration with other curated datasets (e.g., from scientific literature, publicly available databases, pre-publication datasets, and additional datasets from the collaborative project in question) loaded onto the platform to readily perform cross-dataset analyses.

For a data platform to deliver this amount of added value, appropriate resources for personnel and infrastructure must be devoted to its development. Accordingly, funding agencies should make sure that more opportunities exist to support such efforts, and research institutions should create additional incentives to provide further support. Building such an infrastructure is necessary to ensure that data sharing is not only technically feasible, but also that the incentives are properly aligned for all stakeholders. Unfortunately, this foundational

Box 1. An example of data access restrictions and data access trust

For a recent COVID-19 project (COMET), we established the following trust. Over ten research labs agreed to deposit their raw and processed data, including bulk- and single-cell sequencing, cytometry by time of flight (CyTOF), cytokine profiles, and antibody characterization, in near-real-time into the UCSF Data Library for a data trust of open access across 150 researchers who have signed the COMET Data Sharing Agreement. These data are continually curated and aligned to de-identified clinical data that is shared across the project. Progress and insight-level data findings are shared among the COMET team at bi-weekly COMET lab meetings. An excerpt of the Data Sharing Agreement reads as follows:

> *"As members of the collaborative project, you will have access to the project on the data repository, which will host the raw and processed data generated across participating labs. However, our data sharing proposal distinguishes data access from data use. Each data set will be associated with a lab "steward," typically the PI from the lab that generated the data. Despite all investigators on the team having access to the data, there is an expectation of trust: that to make use of a given dataset an interested investigator will contact the steward prior to accessing the data with a specific proposal for data use that the steward can agree to, and that the steward be kept informed of use and progress on the analysis and included as a collaborator. This policy is to strike a balance between promoting collaborative science and respecting the investment the steward lab has put in to generating the data."*

The COMET Data Sharing Agreement includes the following clauses:

- As a lab generating data for the collaborative project, we will facilitate upload of the raw and processed data to the project repository in a timely manner (ideally within 1 day to 2 weeks of generation). This includes ancillary data generated as part of samples acquisition.
- Prior to accessing data on the project for which I am not the steward, I will contact the lab steward to request permission for specified data use and will continue to update them on my use of the data and findings as an involved and respectful collaborator.
- I will present my data insight at the bi-weekly data meeting.
- I agree that should a new investigator request to work with project data as a collaborator, I will direct them to the established process of the request survey system for approval, and that I will confirm they have received approval and signed the data sharing agreement prior to sharing data.
- I will follow COMET's publication and authorship policies.

HOW ITS WORKING: CHALLENGES AND SOLUTIONS

This process has presented specific requirements, triumphs, pitfalls, and solutions—it has enabled a series of new collaborations and much broader data use during this critical moment in the project and course of the pandemic.

Patient privacy and honoring consent

Because of the circumstances of their illness, certain hospitalized patients were enrolled in the study under a waiver of consent, and data could be generated but not shared widely; if patients later declined all associated data needed to be destroyed, but if patients consented all data needed to be made available to the COMET team. This led to complications for data management, inclusion, and sharing in the Data Library.

Solution: we built a restriction system in our database and file server such that if a patient's consent status was "waiver," their records were withheld from search results unless the user had privileged access. If the patient status updated to "consenting," these records and data switched to unrestricted. If patient status updated to "decline," all records and files were automatically removed and queued for deletion.

Equal access to samples, data, and insights

As results and insights developed and were shared during the project, multiple labs could begin working on the same sets of samples and data in real time, leading to conflicts of "ownership" of ideas.

Solution: the project executive committee intervened to resolve conflicts and reorganize priorities and domains, helping researchers re-align and focus on distinct areas, combine efforts, and manage overlap.

Timely data posting and insight-level sharing

Labs varied in timely updates to the data repository with new data, leading to delays in data access and inequitable contributions to the data trust.

Solution: we introduced additional data streams and personnel (project and data management) to confirm and facilitate up-to-date data sharing and lab meeting participation and presentation.

work rarely results in prestigious publications, and it needs special consideration from research institutions in an environment where funding is almost exclusively tied to academic achievement.

Beyond raw and processed data, the sharing of insight-level data would come at the time of regularly held project-wide lab meetings or equivalent—this insight-level data can include additional feature extraction or dimensionality reduction, and/or observed signatures or biology revealed in the data. This creates an atmosphere of openness and inclusion, a forum to integrate insights across investigators, and an opportunity for team members to provide feedback and additional insights of their own. If all attendees are required to agree to the data trust document (see Box 1), there is at the very least a societal norm that governs how data should be protected within the trust.

The next steps for all three types of data will be dissemination to other collaborators, other institutions, and to the public in the form of publications and public data hosting. A key to this process is adaptability and transparency. As new investigators seek access to the trust, they also become explicitly part of the trust and are expected to deposit their own insight-level work with clear understanding of how and by whom it might be accessed. Accordingly, prior to granting access to the project repository, the collaborator must review and sign the data sharing agreement and follow the process outlined in the data trust agreement to access data (contacting stewards for data use and the associated rules). Through this process the entire project improves its access to ‘‘expertly curated’’ data, even beyond what might have initially been conceived. As before, all of this requires monitoring, to minimize unequal sharing and possible data misuse, which would erode trust in the system.

As a use-case, we first implemented and refined this approach in a recent and collaborative COVID-19 study called ‘‘COVID-19 Multi-Phenotyping for Effective Therapies (COMET)’’ at the University of California, San Francisco (UCSF). The agreement and reflections on the process are detailed in Box 1.

Authorship

An important component of this type of data sharing model is an ethos where all investigators make their best efforts toward crediting the hard work and dedication of team members, including those on clinical, biospecimen processing, data analysis and management, and leadership teams. We recommend that project leaders define a set of authorship expectations at the outset of a data sharing agreement that might include a recipe for consortium attribution. Publications that involve data or significant expertise from a steward lab should approach the steward in regard to authorship.

This authorship model is part of a larger need for a culture shift in authorship and credit toward inclusion of all contributing members of a project and requires formally rewriting current authorship guidelines, which historically often exclusively rewarded profound intellectual contribution (The International Committee of Medical Journal Editors, 2020) to the manuscript. This ethos is increasingly inadequate in a world where projects require collaboration between several investigators with different expertise and wherein all contributions are critical but often easily forgotten or underestimated by project leaders. This shift is already happening, helped along considerably by author contribution paragraphs that allow the nature of contributions to become explicit. Although re-writing the authorship code for the age of collaborative science is beyond the scope of this manuscript, we highlight this as a developing issue because authorship is in many cases the primary reward that can be offered to encourage collaboration, and it is an integral part of the advancement system in many institutions worldwide.

Closing and challenges ahead

In summary, the large datasets we now produce have enormous value, but that value is only fully achieved if data can be mined in many different ways by multiple groups. As technology advances, and more data are generated, a majority of existing data, both public and private, are under-analyzed, and therefore, under-utilized. This is particularly wasteful when considering that the amount of data in the public domain dwarfs what can be generated by a single investigator or institution. Incorporating it all—raw and insight level–with other data, years earlier than current sharing requirements enable, will provide enormous value, both to investigators and to science and society as a whole. There are doubtlessly going to be additional refinements to a data sharing trust model. Critically, this model is most easily pioneered among a large number of researchers predominantly from the same institution—therefore there might need to be additional safeguards put in place for pre-publication sharing with researchers across multiple institutions. If there is an expectation of trust, having a new researcher (1) sign the data sharing agreement and (2) be given temporary access to the data sharing with user actions limited and recorded, could provide sufficient protections. Another option for third parties is to limit access to subsets of the data specified by the corresponding steward, without having access to the entire multi-investigator project. Although further developments of this framework will be needed, it would seem timely, particularly on the heels of this century's fifth and largest global health threat, to start refining these models now.

REFERENCES

Bierer, B.E., Crosas, M., and Pierce, H.H. (2017). Data Authorship as an Incentive to Data Sharing. N. Engl. J. Med. *376*, 1684–1687.

Olfson, M., Wall, M.M., and Blanco, C. (2017). Incentivizing Data Sharing and Collaboration in Medical Research-The S-Index. JAMA Psychiatry *74*, 5–6.

Sim, I., Stebbins, M., Bierer, B.E., Butte, A.J., Drazen, J., Dzau, V., Hernandez, A.F., Krumholz, H.M., Lo, B., Munos, B., et al. (2020). Time for NIH to lead on data sharing. Science *367*, 1308–1309.

Birney, E., Hudson, T.J., Green, E.D., Gunter, C., Eddy, S., Rogers, J., Harris, J.R., Ehrlich, S.D., Apweiler, R., et al.; Toronto International Data Release Workshop (2009). Prepublication data sharing. Nature *461*, 168–170.

The International Committee of Medical Journal Editors (2020). http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.