Topics in Companion An Med 37 (2019) 100367

Research Article

Novel Data Sharing Agreement to Accelerate Big Data Translational Research Projects in the One Health Sphere

Joshua Staley, MS[c], Reza Mazloom, MS[d], Paul Lowe, CRA[e], CT Newsum, JD[f], Majid Jaberi-Douraki, PhD[d], Jim Riviere, DVM, PhD[c], Gerald J. Wyckoff, PhD[a,b,*]

Keywords:
data sharing
one health
public-private partnership
data analysis
legal issues
translational research
big data

[a]*Molecular Biology and Biochemistry, University of Missouri – Kansas City School of Biological and Chemical Sciences, Kansas City, MO, USA*

[b]*Division of Pharmacology and Pharmaceutical Sciences, University of Missouri – Kansas City School of Pharmacy, Kansas City, MO, USA*

[c]*Veterinary Biomedical Sciences, Kansas State University, Olathe 22201 West Innovation Drive, Olathe, KS, USA*

[d]*Institute of Computational Comparative Medicine, Department of Anatomy and Physiology, Kansas State University, Manhattan, KS, USA*

[e]*Office of the Vice President for Research, Kansas State University, Manhattan, KS, USA*

[f]*Aratana Therapeutics, Inc., 11400 Tomahawk Creek Pkwy #340, Leawood, KS, USA*

ABSTRACT

When conducting translational research, the ability to share data generated by researchers and clinicians working with for-profit companies is essential, particularly in cases that involve “one health” data (i.e., data that could come from human, animal, or environmental sources). The 1DATA Project, a collaboration between Kansas State University and the University of Missouri, has examined and overcome some of the barriers to sharing this information for “big data” projects. This article discusses some of the obstacles we encountered, and the ways those obstacles can be surmounted via a novel form of Master Sharing Agreement. Developed in collaboration with industry partners, it is presented here as a template for expediting future one health work.

Introduction

Modern data analysis relies on bringing dissimilar sets of data together in a computational resource that enables storage and retrieval of this dissimilar data in a structured way. For researchers working at the intersection between animal and human health, the first challenge is finding unstructured data collection and storage methodologies to sift through office visit/encounter notes, which requires the implementation of natural language processing or other text-mining tools. However, these are generally only useful for organizations that are compiling and mining data from a variety of disparate sources such as marketing agencies, and mining such data using natural language processing or text-mining tools.[1-3] Medical data, on the other hand, tends to be highly structured by the quantitative nature of the field. It is also designed to both protect personal health information, and interface with billing systems, to ensure that insurance and payers are appropriately billed and notified. However, this structure is often antithetical to research applications, as it requires much de-identification and curation to be adapted to a research framework. This is costly in time and resources.

Fundamental research data is collected in a large number of repositories housed by larger organizations including NCBI,[4] EMBL,[5] or DDJB.[6] Many of these repositories are public and also structured along precise lines as defined by a particular research aim or the needs of a funded consortium, leading to a plethora of variably formatted datamarts. This creates significant concerns that need to be addressed to utilize the data, as “best practices” in storage methods change over time. As researchers attempt to harvest stored data, the protocol for accessing that data may antiquated and no longer supported; and documentation may be so poor that interoperability becomes a barrier to research. This requires the researcher to either write a new mining tool, or run and maintain legacy resources to access this data. Regardless, this is not an effective use of time or resources.

The 1DATA Consortium

The 1DATA Project, a collaboration between Kansas State University and the University of Missouri, has examined and overcome some of the barriers to sharing these types of data.[7-9] This project was funded to meet the challenge set forth by BioNexus KC (formerly the Kansas City Area Life Sciences Institute),[10] an advocacy group

[*] Corresponding author. Gerald J. Wyckoff, Molecular Biology and Biochemistry, University of Missouri – Kansas City School of Biological and Chemical Sciences, Spencer Hall room 522, 5009 Rockhill Rd., Kansas City, MO 64110, USA
E-mail address: Wyckoffg@umkc.edu (G.J. Wyckoff).
Abbreviations: DCO, Data Contributing Organization; NLP, natural language processing; SEADS, Structured Environment for Animal Data and Simulation; API, Application Program Interface

http://dx.doi.org/10.1016/j.tcam.2019.100367
1938-9736/© 2019 Elsevier Inc. All rights reserved.

that coordinates across stakeholders in the Kansas City region. They identified data analytic resources, and the nexus of animal and human health, as priority areas for regional development in their “Path to 2025” report. The 1DATA Consortium, made up of researchers from Kansas State University, the University of Missouri Kansas City, and other regional institutions, worked to create a framework for sharing one health research results, and clinical data, in support of collaborative translational research. Much of the initial work that has been carried out so far rests depended on the creation of the SEADS data structure, an enabling framework for the 1DATA project. SEADS (Structured Environment for Animal Data and Simulation) is best thought of as a platform that collects and integrates these data into a coherent structure for retrieval and analysis. In addition, SEADS as well as providing a convenient platform for applications to be developed, also allows access to the database.[11] An Application Program Interface allows developers to build specific “apps” which pull data from specific realms and can combine them to allow complex simulations across never-before-linked data sources. Datasets within the framework can be federated (i.e., separated, encrypted), to allow users access only to portions of the data as determined by the user interface and access control. This allows stakeholders to moderate how their data can be accessed and who can access it, ensuring proprietary data remain as such.

SEADS houses datasets from a wide variety of sources and disciples. These include: phylogenetic data on animal and human pathogens; antimicrobial resistance data on animal and human pathogens; pharmacokinetic/pharmacodynamics; (PK/PD) data; epidemiolog data on animal disease; companion animal health data (e.g., electronic records form veterinary clinics) when available; pathology when available; congruence between human and animal disease data; laboratory animal data; zoonotic disease data.

Some SEADS data sources have been collected based upon strict regulatory guidelines, such as information from the Food and Drug Administration’s Center of Veterinary Medicine (FDACVM) New Animal Drug Applications (NADAs). These data add value to any other data with which they are paired, such as information from companion animal (i.e., cats, dogs, etc.) data sets. From the perspective of a researcher, this would be the first step in an analysis, as the data has a number of properties that make it amendable to study. This includes it containing defined starting and ending points (e.g., minimal incomplete records as cases lost to follow-up would clear by the absence of the ending point); has a denser set of clinical and biomarker monitoring; and, in some cases, these studies include pilot pharmacokinetic studies.

With highly granular data, 1DATA could curate a database of virtual animals (that is, in silico analogs of animal characteristics) integrated from overlapping datasets from real animals in different data sources, which would allow researchers to obtain control datasets of animals useful for their unique research objective, health record requirements, species and/or breed. For example, if a pharmaceutical study was to be completed, and dogs were chosen to be the recipient of a drug that affects the heart; 1DATA could provide the, “control” or baseline, health information. This would allow those evaluating the animals within the study to determine if their dogs are “healthy” before the study begins and determine the effect of the treatment, potentially, without the use of living dog as a control or baseline subsequently reducing the number of dogs required for the study. This meets the requirement for reducing animal use and replacing animals in studies for ethical animal use. In addition, such “control avatars” would be representative of larger and more representative populations of animals compared to small control groups.

However, in setting up this novel datamart geared towards big data analysis projects, it was clear that a major barrier to entry for companies and other entities that wished to share their data using the 1DATA framework was the lack of a comprehensive, structured data agreement that afforded them rights and protection as they shared data with researchers. Thus, we set about an effort to create this Master Data sharing agreement for the 1DATA Consortium. In identifying the issues that were solved with this agreement, and the tools put in place to honor it, the Consortium realized that the process of creating the agreement, some of the ideas behind it, and the agreement itself were useful products that can and should be shared with the one health research community.

Materials and Methods

Identifying Partners: Their Rights, Privileges, and Responsibilities

Identification of stakeholders can be a fraught exercise when sharing data; however, it is an essential first component of any agreement. When developing data agreements, the developers of this agreement need to know who has a stake in the outcome of the research, what that stake is, and what interests might be competing with those interests. For-profit companies that contribute data for data sharing (Data Contributing Organizations [DCO]), might do so for a variety of reasons, however, they will likely expect to realize value from the research performed on it; even if that value is low, a long time away, or considerably removed from the general line of interest of the organization. Basic research is generally tangential to the core interests of a company; translational research, however, will generally produce something tangible within the near- to the mid-term time frame of an agreement, and subsequently more favorable for the for-profit stakeholders.

Creating Clear Expectations: Definition of Terms

Term definitions are crucial for useful data exchange. The specific terms and respective definitions and use cases are generally discussed as being a property of the data exchanged. A clear understanding of these terms was essential for all stakeholders. For example, we sought to create an agreement that would serve both for-profit and not-for-profit companies that generate data, as well as, potentially, nonprofit organizations that broker data from other organizations or sources. We needed to make a clear distinction between the 1DATA team tasked with receiving the information for creation and maintenance – the DCOs, from end-users – the entities that wanted to make use of the cleaned, organized, and de-identified data. We also realized that defining who had access to data, and at what level they could administrate, was a concern not only for the logistics of data security, but also for DCOs who wanted to ensure they had a clear understanding of the exposure of their data. Of concern as well could be researchers accessing the data that wanted to understand its provenance.

Ensuring HIPAA-Compliant “De-Identification” of Data

HIPAA requires “de-identification” of all human health records when they are released for use by those other than the patient or care provider. For this reason, the infrastructure of stored data within the 1DATA framework is HIPAA-compliant, as the act de-identification can be completed when the data is searched and there is no method for looking at the complete, unfiltered dataset. We take the term “de-identified” as a term of research generally meaning that data has had characteristics removed that would allow it to be linked back to a specific individual or small group of individuals. It also refers to data that, by the Master Data Sharing Agreement, should not be subjected to re-identification through force majeure. In the case of human health information, all data is encrypted at all times, and is decrypted when authorized access to that data has been permitted under the terms of the DCOs data sharing agreement, however demographic data from the patient records are not released, Name, address, contact information, while, if deemed necessary for the research a zip code may be released. However, de-identification is dynamic, and in terms of rare disease a zip code may be defined identifiable information and for that reason it would not be released. To ensure this, there is no automatic download or dump of 1Data information, which would allow the unintentional release of identifiable human health information; all such cases are reviewed by member of the 1Data team.

Patient-centered Approach

Data containing age related information provides a good example for discussion of a patient-centered approach to de-identification. A data set for someone interested in gerontology might contain the birth and death dates for a group of individuals, as well as their names and relationships arranged in a family tree or genealogy. If the interest of researchers is to look at genetic heritability of the trait for longevity, an easy way to de-identify the data would be to strip off names and, instead of birth and death dates, utilize absolute ages across the tree. However, assuming birth and death dates are available from a third party, nonaffiliated source (such as cemetery records), the unique combination of marriage and birth-death dates for spouses might allow the simple re-identification of this data. Use of the data would be subject to all parties agreeing to not re-identify data through any means, including the use of third-party data, as defined above.

Data-provider Approach

For many companies that do work in the life sciences, the mere attachment of their name to a record reveals informational context that they (the company) may not have wanted to be made known. For this reason, through the SEADS framework, it is possible to remove any amount of contextual and/or demographic information that the contributing organization sees fit. For example if a company A was working on new class of “Drug X” and shared some of their fundamental data with 1Data and Company B were to search this data set, Company B would find no connection of the fundamental drug and Company A, thereby protecting Company A’s stake in “Drug X.”

Both of the de-identification approaches are implemented simultaneously and respective to the DCOs’ wishes’ and HIPAA compliance. Whether this suffices for all future partners remains an open question, however the system maintains an architecture that should allow even higher levels of auditing and granularity. By agreement, we would need to meet de-identification standards agreeable to the DCOs AND the organization requesting the data.

Managing Data Sharing Risk: What it Means for All Partners

The Principal Investigators from the 1Data project initially approached the issue of managing data sharing risk as synonymous with de-identification of deposited data and removing legal risk from companies that might submit data through indemnification clauses. However, this isn’t entirely the case. A *eureka* moment occurred during the development process when the team realized the need of for-profit companies to ensure that, in cases where data they shared resulted in the creation of something of substantive value, it could be accounted for and potentially captured, allowing shareholders of the company to realize value from the company’s contribution. This led to several modifications to the 1Data database and system (discussed below) that allow tracking and accounting of what data are actually employed in an analysis on a per-project basis. In this way, DCO concerns regarding potential profit sharing and right-of-first-refusal are addressed.

One of the ways to help manage data sharing risk for contributing organizations was to limit the time frame of shared data, as to allow the contributing organization to regain their exclusive right to their data. However, this has the negative consequence of creating a time limit for researchers, who might be continuously working on data analytics when the project agreement ends.

For this purpose, a clause needed to be created in the Master Data Sharing Agreement that allowed some reasonable extension of data use for researchers who were actively engaged in projects. Introduction of such a clause allowed for the creation of a partnership between the researcher and the contributing organization, to allow both parties to benefit from the shared agreement.

Identification of Problems

Potential obstacles to developing an agreement were first recognized in discussions with various researchers and potential stakeholders. A literature review was also performed to ascertain pitfalls that had arisen in other projects. Legal review was sought from counsel across a core group of stakeholder organizations, both University and Private. Last, a face-to-face discussion about the draft agreement allowed identification of issues that might have arisen while other items were solved.

Implementation of an Accounting System for Data Use

As part of the 1Data project, we have asked numerous for-profit entities to share research and clinical data with us as part of a large one health database. The database contains data from public and private sources. Much of the private data would not have value, unless processed with other private or public data. However, 1Data has been asked to manage the risk of our partners when they share data, generally meaning that private companies want to understand their ownership or potential ownership of intellectual property generated in the course of querying the database. Therefore, federating the data (a known property of structured or unstructured databases) is not sufficient for our project.

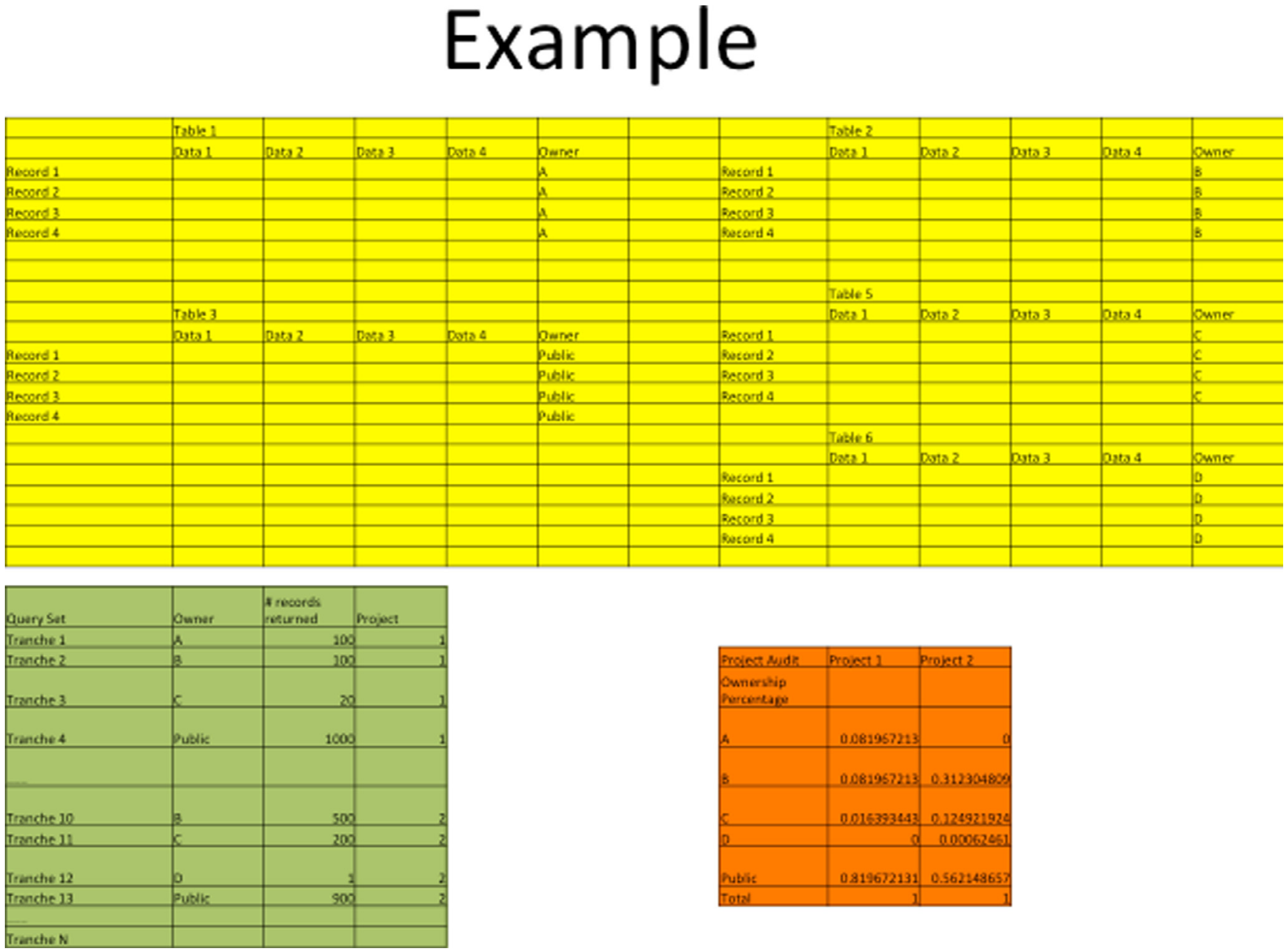
A potential solution is to flag all data at the row-level, to source and owner. This, however, doesn’t address the “ownership” issue for queries involving mixed sets of records with different owners. The problem is solved by recording the ownership percentage and raw numbers of rows (records) returned for each query and in the case of multiple DCOs, this allows auditing to determine, in the end, what percentage of returned rows were from which owner (Fig 2).

Before a search is completed a formal project proposal is started containing information pertaining to data sources (i.e., human, animal, etc.) that are to be searched and potential DCOs if known, at this point a project ID is created and all searches completed are done so using that ID. After the completion of the search, there are 3 components to ensure that project integrity is maintained over time. First are the identification of what data was actually used and the ownership of information used with respect to the overall dataset, as depicted above in Fig 2. The next component is the query storage per project, meaning that all queries that took place stored as labeled with their respective project ID. The third component is the aggregate project audit. This would entail using an exogenous database structure that stores the project ID along with DCOs information. This is currently implemented as a table with restricted access in the main body of the 1Data database (Fig 1). Future implementation of blockchain would allow this record keeping to be automated and ensure integrity of file access and management.

Results and Discussion

Master Data Sharing Agreement

The result of our efforts is the Master Data-sharing Agreement (Supplemental material; link) used for the 1Data project, which we believe will serve as a template for other such agreements. To the best of our knowledge, this is a unique agreement structure. It enables the integration of clinical and nonclinical research data from



Example

	Table 1							Table 2				
	Data 1	Data 2	Data 3	Data 4	Owner			Data 1	Data 2	Data 3	Data 4	Owner
Record 1					A		Record 1					B
Record 2					A		Record 2					B
Record 3					A		Record 3					B
Record 4					A		Record 4					B
								Table 5				
	Table 3							Data 1	Data 2	Data 3	Data 4	Owner
	Data 1	Data 2	Data 3	Data 4	Owner		Record 1					C
Record 1					Public		Record 2					C
Record 2					Public		Record 3					C
Record 3					Public		Record 4					C
Record 4					Public							
								Table 6				
								Data 1	Data 2	Data 3	Data 4	Owner
							Record 1					D
							Record 2					D
							Record 3					D
							Record 4					D

Query Set	Owner	# records returned	Project
Tranche 1	A	100	1
Tranche 2	B	100	1
Tranche 3	C	20	1
Tranche 4	Public	1000	1
……			
Tranche 10	B	500	2
Tranche 11	C	200	2
Tranche 12	D	1	2
Tranche 13	Public	900	2
……			
Tranche N			

Project Audit	Project 1	Project 2
Ownership Percentage		
A	0.081967213	0
B	0.081967213	0.312304809
C	0.016393443	0.124921924
D	0	0.00062461
Public	0.819672131	0.562148657
Total	1	1

Figure 1. Example of 1Data Structure. Top – 5 tables containing data from sources A, B, C, D, public, bottom-left – example of search history containing project id, data source, and number of records, bottom-right – ratio of data sources used per project.

multiple for-profit, not-for-profit, and University partners to a single, implemented framework focused on one health data.

There are 2 tiers at which data can be shared, addressing in particular the concerns of for-profit companies regarding potential profit loss through data sharing.

Tier 1

This is the most conservative case. As depicted in Fig 3 the DCO agrees to share with 1DATA, but marks all of the deposited data as proprietary. Data in this tier is not accessible to anyone except the core investigators, whose task is to curate your data's relationship with the data already deposited from other sources and that your shared data can not be used in research without an agreement first being met as disclosed in the section titled *De-risking data- what it means to all partners.* While submitting data in this way protects your interest in the data, it limits the number of researchers looking at it meaning that any discovery made will likely be by the contributing organization as 1Data resources are limited. This tier is recommended for a DCO that has analytical staff.

Tier 2

This is the case where data is submitted as "nonproprietary, 1Data researchers only." The rule governing this proprietary data remains the same but the DCO has opened a portion of their data source to be partially publically available.

This may lead to the question; what about data that is publicly available or freely shared by a not-for-profit? Regardless of how a DCO chooses to share their data: the public will never have access to submitted data without the consent of the DCO. The workflow as depicted in Fig 4 is as follows: a person navigates to the 1Data website and preforms a query using our Application Program Interface. This only returns a count of results and percentage of how many of those results are proprietary. The person then has the opportunity to submit a project proposal in order to obtain access to the queried

$$\frac{\#\ of\ records\ from\ DCO\ 1}{\#\ of\ records\ in\ Data\ Query} = Percent\ of\ DCO\ 1\ contribution$$

$$\frac{\#\ of\ records\ from\ DCO\ 2}{\#\ of\ records\ in\ Data\ Query} = Percent\ of\ DCO\ 2\ contribution$$

Figure 2. Formulas used to calculate data ownership. Formulas used to calculated ratio of data source used per project, used in Fig 1 example of 1Data Structure.

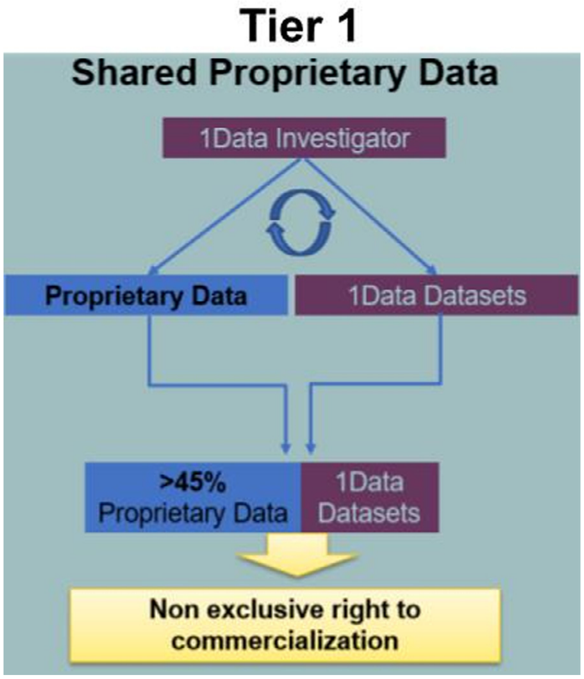


Figure 3. Workflow for Tier 1 data sharing agreement. Workflow used when proprietary data is used in project, in order to determine data ownership.

dataset. This proposal requires approval from the 1Data investigators and the respective DCO(s).

One Health Application

This unique partnership agreement creates an atmosphere where human, animal, and environmental researchers can collaborate in real time with curated, verifiable data. For the past many years the 3 fields have operated independently of each other, and potentially created a sizeable amount of redundant data. The 1Data framework allows all involved disciplines to take advantage of each other's advancements and make some of their own—at potentially a lower cost, and in a shorter amount of time. This creates a unique opportunity for one health data to sit at the forefront of translational and interdisciplinary research.

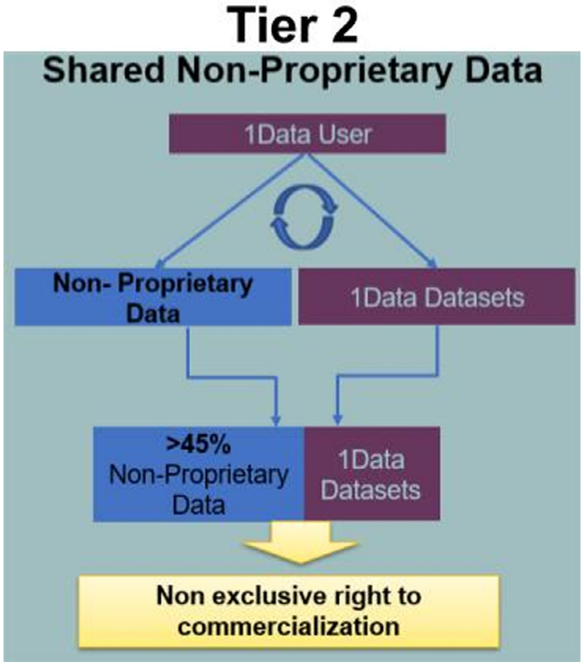

Tier 2
Shared Non-Proprietary Data
1Data User
Non- Proprietary Data
1Data Datasets
>45% Non-Proprietary Data
1Data Datasets
Non exclusive right to commercialization

Figure 4. Workflow for Tier 2 data sharing agreement. Workflow used when nonproprietary data is used in project, in order to determine data ownership.

Translational Research Application

The field of translational research has always presented challenges in time management and data curation. Researchers needed broker deals with at least 2 data sources, and then synthesize that data into a format suitable for further study. This process consumed valuable time and resources and at this point the researcher still had to study the newly formatted data in hopes of finding a novel relationship or characteristic. With the first steps eliminated, the researcher is free to deliver on the promise they made to their funding organization in a timely and efficient manner.

Similarly for researchers seeking insight into human diseases from data in other species (or vice versa), there needs to exist an accessible site where those data exist in a format that can be readily studied. This new convenience would allow for translational medicine to more rapidly and efficiently develop, simultaneously improving the welfare of both humans and animals.

Conclusions

Enabling a data sharing agreement allows for the fruitful use of data that would otherwise have gathered dust within corporate data structures. It has enabled meaningful public-private partnerships and agreements that fuel translational research and allows for the creation of novel intellectual property that otherwise could not have existed because of the admixture of public, private, and research data. The process of creating the Master Data Sharing Agreement for the 1Data project contains lessons that can be applied across future data sharing endeavors.

Acknowledgments

The authors wish to acknowledge Ada Solidar and Jason Wilson for helpful edits and commentary. Funding for portions of this work was provided by BioNexus KC to the 1DATA Consortium, and from the Civic Council of Greater Kansas City.

Submission Declaration

This article has not been published previously in any other forum nor will it be published subsequently in any additional forum.

Contributors

GJW, JR, and MJ conceived of 1DATA and carried out the early conceptualization of the project. JS created the framework for tracking the data ownership. JS and RM worked on the Master Data Sharing agreement in the early phases. PL and CTN worked on the Master Data Sharing agreement and secured permissions. JS and GJW wrote the paper.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.tcam.2019.100367.

References

1. Wei W-Q, et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc JAMIA* **23:**e20–e27, 2016
2. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. *J Am Med Inform Assoc JAMIA* **16:**596–600, 2009
3. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* **19:**1236–1246, 2017
4. National Center for Biotechnology Information (NCBI). *Natl Center Biotechnol Inf* , 1988. Available at: https://www.ncbi.nlm.nih.gov/home/about/mission/ (Accessed: 15th September 2019)
5. Kanz C, et al. The EMBL nucleotide sequence database. *Nucleic Acids Res* **33:** D29–D33, 2005
6. Kodama Y, Mashima J, Kosuge T, Ogasawara O. DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acids Res* **47:**D69–D73, 2019
7. 1Data: Improving the lives of humans and animals. Available at: https://olathe.k-state.edu/research/centers-institutes/1data/. (Accessed: 18th September 2019)
8. Open Source University Collaboration Platform Set to Spark Breakthroughs in Human, Animal Health | Open Health News. Available at: http://www.openhealthnews.com/content/open-source-university-collaboration-platform-set-spark-breakthroughs-human-animal-health. (Accessed: 18th September 2019)
9. University Collaboration Integrates Human and Animal Data. *Bovine Vert*. Available at: https://www.bovinevetonline.com/article/university-collaboration-integrates-human-and-animal-data. (Accessed: 18th September 2019)
10. *'Path to 2025' : Kansas City Region Life Sciences Summary Report*. (2015).
11. UMKC Professor: How businesses can save billions on drug trials. *Kansas City Bus J*. Available at: https://www.bizjournals.com/kansascity/news/2017/04/24/umkc-professor-save-drug-research-clinical-trials.html. (Accessed: 18th September 2019).