



Policy-aware data lakes: a flexible approach to achieve legal interoperability for global research collaborations

Adrian Thorogood 

Centre of Genomics and Policy, McGill University, 740 Dr. Penfield Ave, Suite 5103, Montreal, Quebec H3A0G1, Canada

Corresponding author. E-mail: adrian.thorogood@mcgill.ca

ABSTRACT

A popular model for global scientific repositories is the data commons, which pools or connects many datasets alongside supporting infrastructure. A data commons must establish legally interoperability between datasets to ensure researchers can aggregate and reuse them. This is usually achieved by establishing a shared governance structure. Unfortunately, governance often takes years to negotiate and involves a trade-off between data inclusion and data availability. It can also be difficult for repositories to modify governance structures in response to changing scientific priorities, data sharing practices, or legal frameworks. This problem has been laid bare by the sudden shock of the COVID-19 pandemic. This paper proposes a rapid and flexible strategy for scientific repositories to achieve legal interoperability: the policy-aware data lake. This strategy draws on technical concepts of modularity, metadata, and data lakes. Datasets are treated as independent modules, which can be subject to distinctive legal requirements. Each module must, however, be described using standard legal metadata. This allows legally compatible datasets to be rapidly combined and made available on a just-in-time basis to certain researchers for certain purposes. Global scientific repositories increasingly need such flexibility to manage scientific, organizational, and legal complexity, and to improve their responsiveness to global pandemics.

KEYWORDS: big data, data commons, data governance, data sharing, legal interoperability, modularity

I. INTRODUCTION

Health research involving Big Data approaches, or the training of artificial intelligence and machine learning algorithms (AI/ML), depends on access to numerous data sources.¹ Especially in cases like the current global COVID-19 pandemic, researchers need timely access to numerous data sources from around the globe. Unfortunately, in the absence of dedicated transborder data sharing collaborations and supporting infrastructure, scientific data aggregation—especially in times of crisis—tends to be left to researchers and research organizations. Before any analysis can take place, researchers bear the burden of finding, negotiating access to, and curating fragmented data sources, often at great cost and delay.² Particularly during public health emergencies, rapid international data sharing requires “appropriate infrastructure . . . such as repositories and information technology platforms.”³

A popular model for global repositories is the transborder data commons, a scientific resource that pools or connects many datasets and associated infrastructure. A key challenge for establishing a transborder data commons, however, is defining a “clear governance structure that . . . adheres to national and international ethical and legal requirements.”⁴ Diverse legal requirements may be associated with scientific datasets including copyright or database rights, data privacy laws, health research norms, or contractual requirements to provide data generators with academic credit or intellectual property rights in downstream discoveries. Requirements may differ significantly across national and regional legal frameworks. To bring together datasets from around the world associated with different legal requirements, a transborder data commons must develop a shared governance structure that establishes legal interoperability between datasets. Interoperability generally is characterized by the ability to meaningfully exchange data.⁵ Datasets are legally interoperable where associated legal requirements are sufficiently compatible to allow for their exchange, aggregation, and re-use.⁶

The challenge of establishing a shared governance model is often underestimated. Scientific communities often spend years negotiating governance, delaying data sharing and research. Once established, it may be difficult if not impossible to re-negotiate governance to accommodate valuable new contributions, or to respond to changing circumstances. Furthermore, where datasets are subject to diverse legal requirements, establishing shared governance can also involve important compromise. This process can often involve trade-offs between data inclusivity (what data resources are included

1 Eric E. Schadt, *The Changing Privacy Landscape in the Era of Big Data*, 8 *MOL. SYST. BIOL.* (2012); Misha Benjamin et al., *Towards Standardization of Data Licenses: The Montreal Data License*, ARXIV:1903.12262 [CS, STAT] (2019), <http://arxiv.org/abs/1903.12262> (accessed Jul 29, 2019).

2 Michelle M. Mello et al., *Waiting for Data: Barriers to Executing Data Use Agreements*, 367 *SCIENCE* 150–152 (2020); Katrina Learned et al., *Barriers to Accessing Public Cancer Genomic Data*, 6 *SCI. DATA* 98 (2019).

3 Katherine Littler et al., *Progress in Promoting Data Sharing in Public Health Emergencies*, 95 *BULL. WORLD HEALTH ORGAN.* 243 (2017).

4 *Id.*

5 JOHN PALFREY & URS GASSER, *INTEROP: THE PROMISE AND PERILS OF HIGHLY INTERCONNECTED SYSTEMS* 3 (2012).

6 For a detailed discussion of different definitions of legal interoperability, see Adrian Thorogood, *Towards Legal Interoperability in International Health Research*, November, 2019, <https://tspace.library.utoronto.ca/handle/1807/98411> (accessed Mar 22, 2020) at ch 2.

in the commons) and data availability (how broadly the commons can be accessed and re-used by researchers).

This paper introduces a novel, alternative model for structuring transborder research projects: the policy-aware data lake. This approach is inspired by technical concepts of modularity, metadata, and data lakes. Under this approach, dataset contributors do not have to agree up-front to a single set of legal requirements. Instead, they are free to articulate distinct legal requirements for each contributed dataset and to modify these requirements over time. Data contributors are required, however, to describe the legal requirements applying to their dataset using an agreed-upon menu of legal terms. In other words, all the datasets in a policy-aware data lake must be labeled with standard legal metadata. Here are some illustrative examples of standard terms that could be attached—alone or in combination—to a scientific dataset:

- This dataset can only be used for biomedical research uses;
- This dataset can only be used for non-commercial purposes;
- The data generator must be acknowledged in scientific publications;
- Data users must make no attempt to identify individual participants.

Standard legal metadata makes it possible to determine if two or more datasets within a policy-aware data lake are legally compatible. Compatible datasets can be combined and made available to certain researchers, for certain purposes. As a result, a policy-aware data lake can be rapidly reconfigured into various legally interoperable subsets, each suited to different research purposes or contexts.

This paper proceeds as follows. Part I defines the traditional model for transborder projects—the data commons—and the associated challenge of establishing legal interoperability between datasets. Part II defines an alternative model—policy-aware data lakes—that can optimize scientific data aggregation and re-use, and better handle the growing legal complexity of global research ecosystems. Part III provides examples of existing projects already exhibiting features of policy-aware data lakes. Part IV cautions that this alternative model will only work if certain preconditions are met. Otherwise, policy-aware data lakes risk degenerating into legally fragmented data swamps. Scientific communities should carefully consider the advantages, challenges, and limitations of both models when designing transborder projects.

II. THE DATA COMMONS MODEL AND LEGAL INTEROPERABILITY

To support researchers, entities around the world who collect, generate, and steward health data (such as researchers, healthcare institutions, and governments) can build a transborder data commons. A data commons is a collaborative resource that brings many datasets together and makes them available to researchers. These resources come in many shapes and sizes. Some are centralized, whereas others are loosely connected.⁷ In a centralized data commons, the data are stored within a single infrastructure and governed by a single entity. Centralized commons are not always feasible because they require significant upfront investments and negotiations. There are intermediate

7 Jorge L. Contreras & Jerome H. Reichman, *Sharing by Design: Data and Decentralized Commons*, 350 *SCIENCE* 1312–1314 (2015).

solutions, where data are maintained under the control of distributed entities, who share some common infrastructure. This common infrastructure may involve services to harmonize and conduct quality control of data and metadata, search tools to allow researchers to find datasets of interest, or a common access portal facilitating requests to access multiple resources. Some primarily consist of connected datasets; others also include various data curation services, computing resources, and supporting software tools.⁸

At a minimum, a data commons must be legally interoperable.⁹ This means that the legal requirements associated with different components of the commons must be sufficiently compatible to allow researchers to legally and practically access and use them. Datasets subject to conflicting requirements cannot be legally aggregated and re-used. Where data access processes are fragmented, datasets cannot be practically aggregated and re-used, especially as the size of the commons scales, because transaction costs quickly become excessive.¹⁰ The Research Data Alliance, an international research community organization promoting open sharing and re-use of data, defines legal interoperability in the context of publicly funded research data as “the ability to combine data from two or more sources without conflicts among restrictions imposed by data providers . . . and without having to seek authorization from the data providers on a case-by-case basis.”¹¹ The Research Data Alliance further clarifies three characteristics of legal interoperability:

- *the legal use conditions are clearly and readily determinable for each of the datasets typically through automated means;*
- *the legal use conditions imposed on each dataset allow creation and use of combined or derivative products; and*
- *users may legally access and use each dataset without seeking authorization from data creators on a case-by-case basis assuming that the accumulated conditions of use for each and all of the datasets are met.*¹²

A range of legal requirements may be associated with scientific data, which may stem from copyright or database rights, data privacy laws, health research regulations, or contractual terms.¹³ Launching transborder scientific resources is particularly challenging where they deal with regulated data—such as personal data protected by data privacy laws—which may be subject to multiple, potentially divergent legal definitions and requirements across countries. Jorge Contreras and Jerry Reichman warn that “failure to account for legal and policy issues at the outset of a large transborder data-sharing project can lead to undue resource expenditures and data-sharing structures that may offer fewer benefits than hoped.”¹⁴

8 Robert L. Grossman, *Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data*, TRENDS GENET. (2019).

9 Contreras and Reichman, *supra* note 7.

10 Catherine Doldirina et al., *Legal Approaches for Open Access to Research Data*, LAWARXIV (2018).

11 *Id.* at 8.

12 *Id.* at 8.

13 See generally Thorogood, *supra* note 6.

14 Contreras and Reichman, *supra* note 7 at 1312.

The International Cancer Genome Consortium is a successful example of a transborder data commons, with a central data access process and data use policy for its large collection of cancer datasets. The International Cancer Genome Consortium (now called the 25K Initiative) was a large-scale genomics research initiative aiming to generate and share 25,000 whole genome sequences from 15 jurisdictions to better understand the genetic changes occurring in different forms of cancer.¹⁵ The International Cancer Genome Consortium (ICGC) adopted a tiered access approach, with open access for data unlikely to be linked to other data that could re-identify individual participants, and controlled access for more sensitive data such as raw sequence and genotype files.¹⁶ Sensitive data are accessed through the Data Access Committee Office (DACO) to protect the privacy and reasonable expectations of study participants, uphold scientific community norms of attribution and publication priority, and ensure the impartiality of access decisions.¹⁷ The DACO provides researchers access to the commons in a timely and efficient manner.

Unfortunately, achieving such levels of legal interoperability tends to be a drawn-out and painstaking process.¹⁸ This is especially problematic during global public health emergencies like COVID-19, where timely, international data sharing has taken on a sudden urgency.¹⁹ International scientific communities typically achieve legal interoperability up-front, by negotiating a common legal data governance structure. This consists of shared policies, processes, and safeguards to ensure compliance with the diverse legal requirements associated with contributed datasets. For example, the commons may establish governance to control who can access the commons, for what purposes, and under what conditions. Establishing a shared governance structure often requires years of negotiation within a global scientific community.²⁰ In order to participate in negotiating a shared governance model, potential contributors must first determine what legal requirements apply to their datasets. Articulating these requirements can be challenging even for sophisticated contributors, in light of rapidly evolving data sharing practices and scientific techniques, and associated legal uncertainty. This step can further delay negotiations. Once a governance structure is established, it can be hard to re-negotiate, and a scientific community can find itself locked-in to a static set of rules.

In the face of diverse legal requirements, negotiating a homogenous governance structure can involve significant compromise. A transborder data commons typically has to make a key trade-off between data inclusion and data availability (see Figure 1).

15 International Cancer Genome Consortium, “About Us,” <https://icgc.org/about-us> (accessed January 11, 2020).

16 Yann Joly et al., *Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO)*, 8(7) PLoS COMPUT. BIOL. (2012).

17 *Id.*

18 Anne-Marie Tassé, Emily Kirby & Isabel Fortier, *Developing an Ethical and Legal Interoperability Assessment Process for Retrospective Studies*, 14 BIOPRESERV. BIOBANK. 249–255 (2016).

19 Wellcome, *Sharing Research Data and Findings Relevant to the Novel Coronavirus (COVID-19) Outbreak* (2020), <https://wellcome.ac.uk/coronavirus-covid-19/open-data> (accessed August 8, 2020).

20 See e.g., E. Pisani & S. Botchway, *Sharing Individual Patient and Parasite-level Data Through the WorldWide Antimalarial Resistance Network platform: A Qualitative Case Study*, 2 WELLCOME OPEN RES., 26 (2017) (establishing a transborder data commons for malaria researchers took a “long time to reach agreement on the terms under which contributors would allow their data to be stored in the WWARN database... in significant part because of the legal implications of storing individual patient data.”).

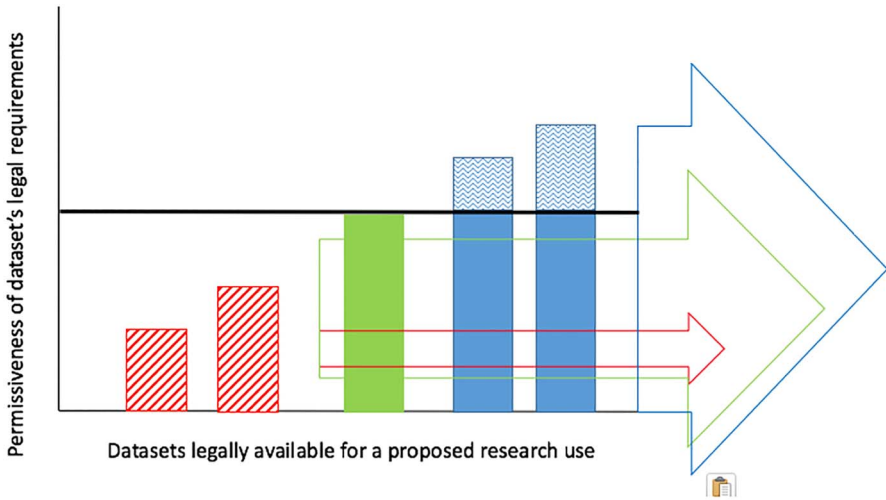


Figure 1. Data Commons: Data Inclusion v.s. Availability. Black line—limit of uses permitted by commons’ governance structure; red arrow—uncontroversial scientific project (3 datasets available); green arrow—somewhat controversial scientific project (3 datasets available); blue arrow—controversial scientific project (0 datasets available); red stripes—datasets subject to restrictive legal requirements that must be excluded from the commons; blue waves—legally permitted uses prohibited by governance structure.

On the one hand, if the data commons establishes a permissive governance model, datasets subject to more restrictive legal requirements are excluded (unless it is possible to re-negotiate these requirements locally). On the other hand, if the data commons establishes a restrictive governance model, more datasets can be included in the commons, but the overall availability of data for research is curtailed. Disagreements within a scientific community over the right balance to strike can often prolong negotiations over governance.

The Human Cell Atlas (HCA) illustrates the trade-off between data availability and inclusivity, primarily because of regional differences in data privacy law standards regarding data identifiability, consent to data processing, and associated safeguards.²¹ The HCA is an international collaboration that aims to produce “a comprehensive reference map of all human cells, meant to serve as a basis for both understanding human health and diagnosing, monitoring, and treating disease. The HCA aims to achieve this by defining all human cells in terms of their distinctive patterns of gene expression, physiological states, developmental trajectories, and location.”²² In addition to reference maps, it will also host individual datasets from cellular biology studies, which now encompass gene expression data, raw RNA sequence data, and associated metadata. One struggle for the HCA has been deciding whether raw RNA sequence data should be made available open access. While data release is legally

21 Mark Phillips et al., *Genomics: Data Sharing Needs an International Code of Conduct*, 578 NATURE 31–33 (2020).

22 Human Cell Atlas, *Ethics Submission Guidance Document* (4 December 2019) https://drive.google.com/file/d/1tHEmgGLj34zf-yCVDg_wAqL_YVsR1acN/view (accessed April 1, 2020).

straightforward for researchers in the USA²³, it is problematic for European projects involving (living) human participants, and subject to the European *General Data Protection Regulation*.²⁴

Once a governance structure is firmly established, a data commons must also maintain legal interoperability over time. This is typically done through complex, lengthy processes of compliance assessment and due diligence by data contributors at the local level.²⁵ Potential new data contributors must assess if the model covers their local legal requirements. At this stage, the governance model is usually take-it-or-leave-it; new contributors have little ability to influence it. This can result in the exclusion of scientifically valuable datasets, perhaps for legal reasons unforeseen at the time the governance structure was established.

These processes raise concerns about murky and unaccountable decision-making. Datasets may be illegally released, which can present legal and reputational risks for the data commons. While legal compliance is primarily the responsibility of contributors, a data commons can provide guidance, compliance assessment tools, or due diligence processes to support responsible contribution. For example, the HCA established “core consent elements” for public (open) sharing of raw RNA sequence data²⁶, and an associated consent template for prospective research studies allowing data to be deposited in the HCA.²⁷ The core consent elements are also being integrated into an assessment tool (forthcoming), to help potential submitters holding already-collected datasets or samples assess if their existing consents permit such data sharing.²⁸

Where contributing a dataset to a commons would breach local legal requirements, a data contributor may, where possible, seek to renegotiate those requirements. Not all legal requirements associated with data are created equal. While some requirements are a matter of public order and must be respected in all circumstances, others arise from private agreements, regulatory authorizations, or individual consents and can potentially be modified or re-negotiated.²⁹ For example, where researchers initially failed to seek participant consent for cross-border transfer or broad re-use of data, they can potentially re-consent participants.³⁰ If such re-consent is impracticable, a research ethics committee may be authorized to waive the consent requirement.³¹ The ability of contributors to modify some legal requirements raises another strategic challenge for a transborder data commons. Should contributors first be expected to negotiate the best possible conditions before the governance model is established? This would allow more datasets to be included in the data commons, or the data commons as a

23 Rudolf I. Amann et al., *Toward Unrestricted Use of Public Genomic Data*, 363 *SCIENCE* 350–352 (2019).

24 *Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC* (General Data Protection Regulation).

25 Tassé, Kirby, and Fortier, *supra* note 18.

26 Human Cell Atlas, *Core Research Consent Elements for Public (Open) Data Sharing* (5 Dec 2019), <https://www.humancellatlas.org/ethics/> (accessed April 1, 2020).

27 Human Cell Atlas, *Main Template Consent Form—Live Donor/Research Participant* (4 Dec 2019), <https://www.humancellatlas.org/ethics/> (accessed April 1, 2020).

28 Human Cell Atlas, *supra* note 21.

29 Thorogood, *supra* note 6 at ch 2.

30 Sabina Gainotti et al., *Improving the Informed Consent Process in International Collaborative Rare Disease Research: Effective Consent For Effective Research*, 24 *EUR. J. HUM. GENET.* 1248–1254 (2016).

31 *Id.*

whole to be made more widely available to researchers. Contributors may not, however, be willing or able to take these resource-intensive steps. Even where they do, this can further delay negotiations of legal data governance. A data commons can mitigate this by providing guidance to contributors on how to renegotiate legal requirements, or by establishing minimum contribution requirements, but ultimately this is a local responsibility.³²

In summary, a transborder data commons needs to establish a shared legal governance structure to achieve legal interoperability between its constituent datasets. This can involve protracted negotiations, as well as compromise between excluding valuable datasets subject to restrictive legal requirements and curtailing the overall availability of the data commons to accommodate such datasets. Further delays can arise as potential data contributors struggle to define local legal requirements and determine if they are legally permitted to release or connect datasets to the data commons. It can be equally difficult to re-negotiate governance over time. The weaknesses of the data commons model—delays, compromise, and inflexibility—are particularly troubling in light of the global COVID-19 pandemic. The pandemic has triggered sudden shifts in scientific priorities, data sharing expectations, and even legal frameworks.³³ It is unclear if existing international scientific resources are able to respond quickly enough to this challenge. Is there a faster and more flexible way to build or repurpose transborder scientific resources, without sacrificing legal compliance?

III. POLICY-AWARE DATA LAKES

Borrowing on technology and data science concepts, this article proposes a promising alternative to a data commons model for transborder scientific resources: the **policy-aware data lake**. Recall that a transborder data commons achieves legal interoperability upfront by establishing a common legal data governance structure, achieved through extensive negotiation and compliance assessment processes. A policy-aware data lake, by contrast, is characterized by a modular approach to achieving legal interoperability (see [Table 1](#) for a comparison). The flexibility inherent in this modular approach reduces the need for prolonged, upfront negotiations over legal data governance before datasets can start to be pooled or otherwise connected, and in turn made available to researchers. Speed, flexibility, and scalability are highly desirable when seeking to building transborder resources, especially in response to global public health emergencies.

The concept of a policy-aware data lake draws on three technology and data science concepts:

Modularity. Modularity is “the degree to which a system’s components may be separated and recombined, often with the benefit of flexibility and variety in use.”³⁴ The concept is used in the design of complex systems—from industries to software—by breaking down the system into modules, which are “units in a larger system that are

32 Global Alliance for Genomics and Health, *Consent Policy* (2019), <https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/> (accessed April 1, 2020).

33 Global Alliance for Genomics and Health Regulatory and Ethics Work Stream, *Responsible Data Sharing to Respond to the COVID-19 Pandemic: Ethical and Legal Considerations* (v 3.0), <https://www.ga4gh.org/covid-19/> (accessed August 8, 2020).

34 Wikipedia, *Modularity* (2020), <https://en.wikipedia.org/w/index.php?title=Modularity&oldid=939961252> (accessed Feb 17, 2020).

Table 1. Legal Interoperability Models for Transborder Research Projects

	Data Commons	Policy-Aware Data Lake
Definition	A research resource that pools or connects datasets together to make them available to researchers	A research resource that pools or connects datasets together. Each dataset can have distinct legal data governance. Researchers can aggregate and re-use all datasets legally compatible with their context and purpose
Process to Achieve Legal Interoperability	A scientific community establishes a shared legal data governance structure through up-front negotiation	Data contributors describe each dataset using explicit, standard, and accurate legal metadata
Dataset Inclusion Criteria	The shared legal data governance structure must respect the legal requirements associated with dataset	The legal requirements associated with the dataset provide a reasonable likelihood of aggregation and re-use
Legal Availability of Data for Re-use	All datasets within the commons are available for research uses that respect the legal requirements of the most-restrictive dataset included in the commons	All datasets legally available for a proposed research use
Ability to Respond to Changing Legal Requirements	Limited. The common legal data governance structure must be re-negotiated	Data contributors are free at any time to update their legal metadata
Data flows across borders	Yes, by definition	Yes, though some datasets may be subject to distinct data localization rules
Shared infrastructure	Spectrum from centralized to fully distributed	Spectrum from centralized to fully distributed

structurally independent of one another, but work together.”³⁵ Modules have freedom with respect to their internal design as long as they respect certain design rules which allow them to interact with other modules. Modularity deals with complexity in two ways: (i) through abstraction, which hides the internal complexity of modules from system designers, and (ii) through interfaces, which define how different modules interact.³⁶ In a policy-aware data lake, different data contributions are treated as independent modules, seeing as they may be subject to different legal requirements. Each module is permitted to articulate or change its own legal governance.

Metadata. Metadata is simply data describing other data. The popular FAIR principles for data science highlight the importance of metadata for enabling the comparison, aggregation, and re-use of datasets. The FAIRness principles establish a standard for ensuring data are findable, access, interoperable, and re-useable.³⁷ The FAIRness of data is intimately related to the quality of the associated scientific metadata. Metadata is also a kind of interface, which allows researchers to determine what datasets can be compared or combined. High-quality scientific metadata is rich, standard, and accurate. In a policy-aware data lake, data contributions must be described by high-quality *legal* metadata. Legal metadata is data describing the legal requirements associated with a particular dataset.

Data Lakes. Data lakes are an alternative data management model to an organizational data warehouse or a community data commons.³⁸ A data commons typically involves intensive efforts to curate and harmonize datasets at the outset to ensure they are scientifically interoperable. The scientific quality and interoperability of datasets is addressed at data ingress, i.e., when data are contributed. Data lakes, by contrast, allow all of an organization’s or scientific community’s datasets to be deposited at the outset, while deferring scientific data curation and harmonization until a later point (e.g., at the time data are accessed, or afterwards).³⁹ Part of this philosophy is admitting, especially in the era of Big Data, that it is impossible to predict in advance what data assets will be valuable in the future. Big Data is characterized by the four Vs of volume, variety, velocity and uncertain veracity.⁴⁰ In particular, velocity and uncertain veracity encourage a data lake approach, where scientific evaluation and curation are left until the time where there is a clear research use case for particular data assets. Trying to structure and curate data in advance can be inefficient (as some data will not be used) and ineffective (as the structure may not be fit for the eventual purpose). Similarly, a policy-aware data lake allows datasets to be contributed or connected regardless of differences in their associated legal requirements. Like a scientific data lake, a policy-aware data lake does not try and predict in advance what datasets will be valuable to what researchers for what purposes. Whether or not various combinations of datasets

35 Carliss Young Baldwin & Kim B. Clark, DESIGN RULES: THE POWER OF MODULARITY 63 (2000).

36 *Id.* at 63.

37 Mark D. Wilkinson et al., *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, 3 SCIENTIFIC DATA (2016).

38 Robert L. Grossman, *Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data*, ARXIV:1809.01699 [CS, Q-BIO] (2018), <http://arxiv.org/abs/1809.01699> (accessed Dec 3, 2019).

39 *Id.*

40 IBM BIG DATA & ANALYTICS HUB, THE FOUR V'S OF BIG DATA (2013).

are legally compatible and available to be accessed is determined at a later point (e.g., at the time a specific research access request is made).

The legal complexity presented by transborder projects can be addressed through analogous strategies. A policy-aware data lake is defined by three essential characteristics:

Modularity/Flexibility: scientifically relevant datasets can be contributed to a policy-aware data lake, even if they are subject to quite different legal requirements.⁴¹ This addresses the data commons problem of data inclusion (see Figure 2). This flexibility is essential in the era of Big Data, where researchers seek to link diverse datasets together, from diverse sources, subject to diverse legal requirements. By providing this flexibility and independence to data contributors, a policy-aware data lake can be launched and scaled up quickly. Modularity therefore alleviates problems of prolonged negotiations and compliance assessments encountered by the data commons model. The importance of modularity is that it helps to optimize the legal availability of datasets for research (see Figure 2). Not all researchers will need to use all datasets for all research purposes. A data commons defines in advance the extent to which its entire catalogue of data will be legally available, which may be inefficient and ineffective. In a policy-aware data lake, different modules (datasets) can be combined into various different subsets. Each subset of modules is legally interoperable, meaning the legal requirements associated with each dataset in the subset are sufficiently compatible to permit re-use for certain research purposes. Some subsets will have fewer datasets but more permissive requirements (see Figure 2—Blue Arrow). Other subsets will have many datasets, but more restrictive requirements (see Figure 2—Red Arrow). Essentially, the modules of a policy-aware data lake can be reconfigured into various, smaller data commons, which can each be legally made available to certain researchers for certain purposes.

Modularity also enables legal data governance to evolve “along with the data sharing zeitgeist.”⁴² Recall that a transborder data commons cannot easily change its governance structure once established. Consider for example the WorldWide Antimalarial Resistance Network (WWARN), which developed a transborder data commons to support data sharing and improve the tracking of drug-resistant malaria. This data commons demonstrated that “[o]nce set, rules of engagement can be hard to change.”⁴³ The commons might have been able to adopt more permissive legal governance over time, as trust developed within the scientific community, but found it was locked in to its original governance structure. Legal requirements associated with data may also change suddenly, such as in the case of the current COVID-19 pandemic. During this public health emergency, data sharing policy makers have urgently insisted on rapid sharing of results and relatively unconditional data release to the international scientific community.⁴⁴ Furthermore, during public health emergencies, data stewards may be granted exceptional, broad legal authorization to share and use regulated data

41 A minimal standard of legal availability could be established, such as a reasonable prospect of being legally aggregated and re-used.

42 Pisani and Botchway, *supra* note 20 at 29.

43 *Id.* at 29.

44 Wellcome, *supra* note 19.

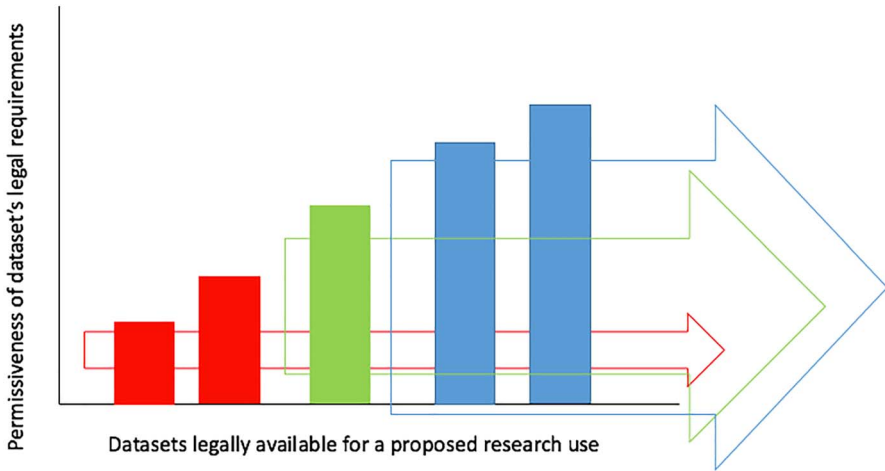


Figure 2. Policy-Aware Data Lake: Data Inclusion v.s. Availability. Figure Legend: Red arrow—uncontroversial scientific project (5 datasets available); green arrow—somewhat controversial scientific project (3 datasets available); blue arrow—controversial scientific project (2 datasets available).

for research, for example personal data governed by data privacy laws.⁴⁵ These policy and legal changes may also be temporary, suddenly reverting to the status quo once the emergency has passed. A transborder data commons is ill-equipped to adapt its shared legal data governance structure to change, whether slow or sudden. Policy-aware data lakes, by contrast, offer greater flexibility, providing individual data contributors ongoing freedom to unilaterally modify the legal governance associated with their datasets.

Legal Metadata. Datasets subject to fragmented legal requirements and access processes cannot be meaningfully aggregated and re-used. How does a policy-aware data lake overcome this problem? Because a policy-aware data lake does not establish legal interoperability between datasets at the outset, it must be able to do so rapidly at a later point in time. This is where the policy-awareness aspect of the data lake comes into play: each data contribution must be associated with high-quality legal metadata. Legal metadata is simply data that describes the legal requirements associated with a dataset. For example, a policy-aware data lake may allow data contributors to define the scope of research purposes for which a dataset is legally permitted to be used. Contributors may be provided with a menu of options to choose from, to allow them to comply with their local legal requirements. Options might include any scientific research; health or biomedical research; diabetes-specific research only; and/or non-commercial research only.

In terms of modularity, one can think of legal metadata as an interface describing how different datasets (modules) interact, i.e., how they can be used, compared, or combined. Legal metadata is also a form of abstraction, which hides the complexity

45 Office of the Privacy Commissioner of Canada, *Privacy and the COVID-19 Outbreak* (2020), https://priv.gc.ca/en/privacy-topics/health-genetic-and-other-body-information/health-emergencies/gd_covid_202003/ (accessed Apr 5, 2020).

of the legal context from which legal requirements arise. A design rule of a policy-aware data lake is that each module must be described with high-quality legal metadata. In other words, datasets must be explicitly and accurately labeled with legal requirements selected from a standard menu. While contributors have significant freedom over *what* legal requirements can apply to datasets, they are constrained with regards to *how* they express those requirements.

High-quality legal metadata are necessary to enable rapid determinations of what datasets are legally available for a particular research purpose. High-quality legal metadata is explicit, standard, and accurate. Contributors must clearly state the legal requirements associated with their datasets (transparency over data sharing conditions is in any case increasingly required by regulators, and expected by research participants). The legal requirements must be described using a standard, commonly understood terminology. Examples of such standards are discussed below. Contributors must also ensure the descriptions are accurate so the legal metadata can be actionable, supporting rapid decisions about providing researchers access to data.⁴⁶ Ideally, machine-readable metadata can help to automate processes of determining what datasets are legally available for particular research purposes. Admittedly, establishing high-quality legal metadata is far from trivial. Most of the key challenges to realizing transborder policy-aware data lakes, discussed below, are metadata challenges.

Rapid Legal Interoperability Assessment: a policy-aware data lake does not establish legal interoperability when datasets are initially contributed. It therefore needs a rapid mechanism to determine what modules (i.e., datasets) can be legally combined and made available to certain researchers for certain purposes. This determination would need to be made “just-in-time” in response to a data access request for a specific research purpose. Practically speaking, a policy-aware data lake might function as follows. First, a researcher submits a request to access all scientifically relevant datasets that are legally available for his or her context and purposes. The data lake then compares the nature of the access request against the legal metadata of its constituent datasets. Finally, the policy-aware data lake aggregates the legally available datasets and provides the researcher access. While a data commons establishes legal interoperability at the ingress phase (deposit of datasets), a data lake establishes legal interoperability at the egress stage (provision of access to datasets). This can only occur if each dataset is described by high-quality legal metadata. Machine-readable metadata may also be desirable to carry out this matching process quickly, as the number of datasets and diversity of legal requirements scales.

Policy-aware data lakes are modular, encouraging inclusion of diverse datasets from around the world, even if they are subject to different legal requirements. High-quality legal metadata is a kind of legal interface, describing how different datasets can be legally combined and re-used. A policy-aware data lake can be reconfigured into various, legally interoperable subsets in real-time. This ensures the legal availability of data is optimized for different research contexts and purposes. By providing flexibility, and

46 While I focus on using legal metadata for providing access to data, it can have other benefits, such as (i) allowing researchers to search for and find legally available data; (ii) assisting regulatory bodies (such as research ethics committees and data access committees) to determine if a dataset is legally available for a proposed research use; and (iii) use as an accountability tool, as it is permanently associated with the dataset, facilitating communication of legal requirements and auditing of researcher compliance.

by reducing the trade-off between data inclusion and availability, policy-aware data lakes can avoid the delays, compromise, and governance lock-in encountered by the data commons model. I now turn to some existing examples of policy-aware data lakes, before discussing their challenges and limitations.

IV. EXAMPLES OF SCIENTIFIC RESOURCES RESEMBLING POLICY-AWARE DATA LAKES

There are already scientific resources exhibiting some characteristics of policy-aware data lakes. The US dbGaP is a central repository for genomic and health-related data from studies funded by the National Institutes of Health.⁴⁷ While the goal of the repository is to maximize the sharing and broad re-use of datasets among the genomics community, the repository has long recognized that datasets from certain research projects may come with distinctive data use limitations.⁴⁸ When researchers deposit datasets into dbGaP, they are asked to specify any data use limitations according to a standard list.⁴⁹ These data use limitations are then enforced by dbGaP's data access committees when researchers seek access to data. The Broad Institute is piloting a software system called DUOS that can assist dbGaP's data access committees to determine if data access requests comply with data use limitations for requested datasets.⁵⁰ The DUOS system is a software system based on the Data Use Ontology, a standard ontology of data use terms maintained by the Global Alliance for Genomics and Health (GA4GH).⁵¹ dbGaP reflects many of the characteristics of a policy-aware data lake. It accepts datasets subject to distinctive data use limitations. Standard data use metadata for each dataset are captured during the submission process. This data use metadata is reliable, as contributors know that the metadata will be acted on by a data access committee. Also, with the implementation of the DUOS system, dbGaP will be able to automatically determine what subsets of its data resources are ethically "available" for a particular access request.

A policy-aware data lake that crosses borders is a different beast altogether. Such a data lake requires a globally accepted legal metadata standard, able to express legal requirements emanating from diverse legal frameworks. The GA4GH Data Use Ontology is an emerging global standard that can be mapped to at least some legal requirements.⁵² One transborder project resembling a policy-aware data lake is euCanShare. This project "is a joint EU-Canada project to establish a cross-border data sharing and multi-cohort cardiovascular research platform . . . [that] integrates more than 35 Canadian and European cohorts making up over 1 million records . . ." ⁵³ The project is seeking to establish a governance structure that respects open science tenets but also complies with diverse applicable legal frameworks. One of its proposals is to develop

47 NCBI dbGaP, *Home*, <https://www.ncbi.nlm.nih.gov/gap/> (accessed Apr 6, 2020).

48 Dina N. Paltoo et al., *Data Use Under the NIH GWAS Data Sharing Policy and Future Directions*, 46 NAT. GENET. 934 (2014).

49 National Institutes of Health, *Standard Data Use Limitations*, https://osp.od.nih.gov/wp-content/uploads/standard_data_use_limitations.pdf (accessed Apr 1, 2020).

50 Broad Institute, *Broad Data Use Oversight System*, <https://duos.broadinstitute.org/> (accessed Apr 2, 2020).

51 *Id.*

52 Global Alliance for Genomics and Health, *Data Use Ontology, Ontology for Consent Codes and Data Use Requirements*, (2020), <https://github.com/EBISPOT/DUO> (accessed Mar 23, 2020).

53 euCanSHare, *euCanSHare*, <http://www.eucanshare.eu/> (accessed Mar 9, 2020).

a data access portal to facilitate access to the project's multiple research resources. This portal is built on the existing infrastructure of the European Genome-Phenome Archive (EGA), which allows research projects to deposit and manage their data using central infrastructure.⁵⁴ The contributing research projects would remain responsible for establishing and enforcing their own access policies, through the establishment of a local data access committee, though the EGA can provide centralized infrastructure for granting and managing dataset access credentials.⁵⁵ One of the deliverables of this project is to code the consent forms and associated documents of contributing projects into machine-readable data use profiles. While the access portal of euCanSHare will permit contributing projects to establish their own local data access policies and procedures, it still aims to ensure data use terms are expressed in a standard, machine-readable format.⁵⁶ The metadata framework euCanSHare is using to represent legal metadata is the Automatable Discovery and Access Matrix.⁵⁷ This metadata model was also developed under the auspices of the GA4GH, and is similar to the GA4GH Data Use Ontology. These data use profiles can then be fed into a search engine, allowing researchers to find datasets across the consortium that are ethically and legally available for their research purposes and context. The consortium will also explore the extent to which a computable approach can improve the ability to both automate and document researcher access to multiple datasets.⁵⁸

These examples reveal concrete differences between how one constructs a data commons and how one constructs a policy-aware data lake. A data commons begins with upfront negotiations over a shared legal governance structure, based on an *ex ante* vision of the resource's purpose. Datasets can only be contributed to the data commons if they comply with the existing governance structure. Researchers then typically request access to the data commons as a whole. A policy-aware data lake, by contrast, does not make upfront decisions about what datasets can be included or excluded. Instead, it begins by simply mapping the legal requirements associated with each dataset to a standard menu of terms. Researchers then request access to the subset of datasets that are legally available for their proposed purpose.

A data commons may be a better model for transborder projects where datasets are subject to relatively homogenous legal requirements, and where a scientific community has a clear, shared vision for how the resource will be used. A policy-aware data lake may be preferred where datasets are subject to more diverse legal requirements, and where the purposes of a transborder resource are likely to evolve over time. While policy-aware data lakes offer potential advantages of flexibility, speed, and scalability, they also come with new challenges.

54 European Genome-phenome Archive, *Home*, <https://www.ebi.ac.uk/ega/home> (accessed Apr 6, 2020).

55 European Genome-phenome Archive, *Browse DACS*, <https://ega-archive.org/dacs> (accessed Apr 6, 2020).

56 Claudia Vasallo, *euCanSHare. Deliverable D3.1—Data Management Plan* (2019), <https://zenodo.org/record/3571022#.XfIQ8uhKiUk> (accessed Mar 9, 2020).

57 J. Patrick Woolley et al., *Responsible Sharing of Biomedical Data and Biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M)*, 3 NPJ GENOM. MED. 17 (2018).

58 Vasallo, *supra* note 56.

V. CHALLENGES AND LIMITATIONS OF POLICY-AWARE DATA LAKES

If not designed and implemented carefully, policy-aware data lakes can degrade into legally fragmented **data swamps**. A data swamp is a collection of superficially pooled or connected data resources providing a mere aura of aggregation, but little meaningful opportunity for researchers to aggregate, access, and re-use data. This is analogous to the scientific data management context where data lakes, due to a lack of upfront data curation to ensure scientific quality and interoperability, end up being scientifically useless.⁵⁹ Policy-aware data lakes may be susceptible to such degradation, because they do not establish legal interoperability upfront. The majority of data sets may end up being subject to conflicting or highly restrictive legal requirements. A scientific community may put significant effort into pooling or connecting datasets before it becomes clear that there is no real prospect of legally compliant data aggregation and re-use.

To avoid this problem, policy-aware data lakes could incorporate some of the harmonization processes used to establish legal interoperability for a data commons (discussed above). Scientific communities could, for example, negotiate minimum legal availability standards for contributions to ensure included datasets have a reasonable prospect of being aggregated with other datasets and re-used. Contributors could also be encouraged to establish the most permissive legal profile for data possible before tagging datasets with legal metadata. Indeed, contributors may be able to modify some legal requirements, by re-negotiating agreements, consents, or approvals. Contributors could also be encouraged to avoid excessively conservative interpretations of local legal requirements.⁶⁰ These harmonization processes would, however, be far more lightweight than in the data commons context. Admittedly, there are many problems of legal interoperability that cannot simply be resolved by negotiation between private parties; the success of policy-aware data lakes will also depend on the continued evolution of background data sharing policy⁶¹ and international legislative harmonization.⁶²

The frictionless vision presented here of a modular and scalable transborder data sharing resource depends on the existence of a global standard for expressing legal metadata. Establishing such a standard is complicated both conceptually and procedurally. Conceptually, a legal metadata standard would, at a minimum, consist of a controlled vocabulary of terms describing different legal permissions, restrictions, and requirements that may apply to the release and use of scientific data.⁶³ A slightly more complex legal metadata standard could be in the form of an ontology, which is a structured and hierarchical terminology. The GA4GH Data Use Ontology, for example, includes hierarchies of data use terms such as general research use, health/medi-

59 Rihan Hai, Sandra Geisler & Christoph Quix, *Constance: An Intelligent Data Lake System*, in PROCEEDINGS OF THE 2016 INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA 2097–2100 (2016) at Abstract.

60 See e.g., National Institutes of Health, *Points to Consider in Developing Effective Data Use Statements*, https://osp.od.nih.gov/wp-content/uploads/NIH_PTC_in_Developing_DUL_Statements.pdf (accessed Apr 2, 2020).

61 Jorge L. Contreras & Bartha M. Knoppers, *The Genomic Commons*, 19 ANN. REV. GENOM. HUM. GENET. (2018).

62 Rolf H. Weber, *Legal Interoperability as a Tool for Combatting Fragmentation*, GLOB. COMM. INT. GOVERN. (2014), <https://www.cigionline.org/publications/legal-interoperability-tool-combatting-fragmentation> (accessed Mar 7, 2019).

63 See e.g., National Institutes of Health, *supra* note 46.

cal/biomedical use, and disease-specific use. Legal requirements can, however, involve complex interdependencies. For example, under the European *General Data Protection Regulation*, personal data may be transferred to countries outside of Europe, but in some cases this may only be permitted on the condition that standard contractual clauses are in place.⁶⁴ To address this, a legal metadata standard may need to go further and include embedded logic to capture interdependencies.⁶⁵ An example of such a standard including a basic logic is the HL7 FHIR Consent standard, which allows healthcare providers to capture, communicate, and enforce a patient's privacy consent directive, an agreement determining how the patient's health information will be accessed, used, and disclosed.⁶⁶

Legitimate consensus-building processes are also needed to establish a global legal metadata standard. Admittedly, this may require higher levels of international collaboration than negotiating a shared legal data governance structure for a data commons. Developing any international technical standard requires high-levels of coordination and collaboration, effort that also needs to be sustained over time to maintain the standard.⁶⁷ Doing this for a standard that expresses legal concepts is perhaps even more difficult. Not only would the standard need to capture legal requirements derived from diverse legal systems, it would also require a process to establish consensus on legitimate definitions or categories of normative relevance. A collaborative effort beyond any given scientific collaboration may be needed, perhaps by a consortium of global scientific consortia or a formal standards development organization. Moreover, in order to restrain a proliferation of competing standards, this entity may need to be an authoritative organization, as was needed to establish standard licenses for creative works (Creative Commons) and open-source software (Open Source Initiative).⁶⁸ Once a legal metadata standard is established, however, it can be quickly re-purposed by various scientific communities. Ultimately, the aim is to establish a foundational standard that becomes invisible infrastructure, which scientific communities can take for granted.⁶⁹

VI. CONCLUSION

Modularity will become an important governance principle for the era of Big Data, in order to handle the growing scientific, organizational, and legal complexity of international research systems. Scientific opportunities to combine data across jurisdictions, sectors, and contexts now far outpace the ability of transborder projects to negotiate a shared legal data governance structure. These opportunities will also continue to outpace lawmakers' efforts to harmonize legal requirements across countries, as well as data stewards' efforts to renegotiate agreements, consents, and other private order sources of legal requirements associated with data. The COVID-19 pandemic has

64 *Supra* note 23, art 46.

65 Woolley et al., *supra* note 57.

66 HL7, *Consent—FHIR v4.0.1*, <https://www.hl7.org/fhir/consent.html> (accessed Jan 15, 2020).

67 Michel Girard, *Big Data Analytics Need Standards to Thrive: What Standards Are and Why They Matter* 209 CIGI Papers (2019).

68 Creative Commons, *When We Share, Everyone Wins*, <https://creativecommons.org/> (accessed Apr 6, 2020); Open Source Initiative, *News*, <https://opensource.org/> (accessed Jun 3, 2019).

69 Girard, *supra* note 67 at 2.

brought home the pressing need for infrastructure that supports international research collaboration. The data commons model will remain an important, perhaps ideal, model for transborder data sharing. Policy-aware data lakes present a promising new alternative for projects dealing with significant legal heterogeneity, or prioritizing speed and flexibility. Pilots are needed to identify this model's limits, and to demonstrate its potential to optimize responsible data aggregation and re-use.

ACKNOWLEDGEMENTS

I would like to thank the funding support of Genome Canada, Genome Quebec, and the Canadian Institutes of Health Research. I would also like to thank Professor Bartha Knoppers and Alexander Bernier for their invaluable feedback on earlier drafts of this paper.