DATA ACCESS

Sharing by design: Data and decentralized commons

Overcoming legal and policy obstacles

By Jorge L. Contreras[1]* and Jerome H. Reichman[2]

Ambitious international data-sharing initiatives have existed for years in fields such as genomics, earth science, and astronomy. But to realize the promise of large-scale sharing of scientific data, intellectual property (IP), data privacy, national security, and other legal and policy obstacles must be overcome (1). While these issues have attracted significant attention in the corporate world, they have been less appreciated in academic and governmental settings, where solving issues of legal interoperability among data pools in different jurisdictions has taken a back seat to addressing technical challenges. Yet failing 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. Drawing on our experience with the Belmont Forum, a multinational earth change research program, we propose a framework to help planners create data-sharing arrangements with a focus on critical early-stage design decisions including options for legal interoperability.

A rich literature beginning with the work of Ostrom (2) addresses the organization and governance of common pool resources shared by communities of users in contexts ranging from the global environment to communal living spaces. More recent work has expanded these principles to knowledge commons: collections of intangible resources, such as digital libraries, scholarly publications, and scientific data (3). Responding to calls for increased international scientific collaboration , several expert bodies have developed high-level principles for transborder data sharing (4–6). Although these efforts lay the groundwork for broad data-pooling initiatives, critical design decisions must be made before addressing larger issues of governance and operation.

A SPECTRUM OF CENTRALIZATION. Although little empirical research exists on commons structures for data sharing and related costs, we have observed four basic structural models for scientific data pools along a continuum ranging from the most to the least centralized (see the table).

(i) *fully centralized*: all data are aggregated in a single, centrally managed repository;

(ii) *intermediate distributed*: repositories are distributed and separately maintained, sometimes across national borders, but may be interconnected by a central access portal, may share other technical service components, and may utilize a common data-exchange format [sometimes referred to as a federated database system (7)];

(iii) *fully distributed*: repositories are maintained locally and are not technically integrated, but share a common legal and policy framework that allows access on uniform terms and conditions (legal interoperability);

(iv) *noncommons*: repositories are largely disaggregated and lack technical and legal interoperability and, at most, may share a common index.

It is not surprising that centralized data repositories with curation, analytics, and quality control can significantly enhance the value of the data they contain [e.g., the GenBank repository of DNA and RNA sequence data (8)]. Centralized structures, however, come at a cost and may be impractical in many transborder collaborations because of political, legal, and organizational issues. But the alternative to a fully centralized commons need not be a noncommons. The shortfalls of noncommons models include incompatible data formats, inability to search across data sets, underutilization of data resources, individualized and inefficient access requirements, and difficulties moving data across national boundaries. Distributed commons structures, however, offer a meaningful subset of benefits with lower cost and resource commitments than fully centralized models.

For example, an online portal through which researchers can access multiple independent repositories may feel like a centralized commons to users, but it avoids the cost and governance overhead of a centralized repository [e.g., the Global Earth Observation System of Systems (GEOSS)]. Portal-based structures may also make it easier for a central administrator to provide users with value-added services and aggregated statistics [e.g., the World Data Center for Microorganisms (WDCM) (9)], and allow users to query multiple repositories simultaneously and more easily combine and analyze multiple data sets (7).

Even if resources do not exist to link repositories technically, there are advantages to fostering legal interoperability among distributed repositories (10). To achieve this across jurisdictions, rules for data usage and access must be compatible with each other, must comply with laws and regulations of the relevant jurisdictions, and must address rights of ownership and control granted to data generators (11). If achieved, legal interoperability can enable researchers to access and use data across multiple repositories without seeking authorization on a case-by-case basis, which increases the likelihood that more data will be put to productive use.

Perhaps the most straightforward path to legal interoperability is simply contributing data to the public domain and waiving all future rights to control it (11). This approach has been advocated by more than 250 organizations that have endorsed the 2010 Panton Principles for open data in science (12). Alternatively, researchers who wish to receive attribution credit for their contributions, but are otherwise willing to relinquish control over them, have released data under standardized Creative Commons (CC) licenses that have been widely used for other online content, including open-source code software, music, and photographs.

Despite the simplicity and appeal of these approaches, they are not always feasible. Data will often remain subject to legal regulation that, for instance, explicitly or implicitly reveal personally identifiable information, were obtained from human research subjects, relate to sensitive technologies, or disclose infrastructural details. Wilbanks and others, recognizing these requirements, have called for new models of informed consent and privacy protection to facilitate broad, socially beneficial sharing of at least some categories of such data (13).

DESIGN CONSIDERATIONS. If a collaborative research project has sufficient resources to create a centralized data repository with accompanying infrastructure and staffing (potentially millions of dollars upfront and thereafter for fully staffed and curated repositories), important benefits can be achieved. In most cases, however, this level of funding will not be available and a distributed data commons could be a desirable alternative. We found, in our experience with the Belmont Forum, that the project's leadership gave substantial weight to early aspirational statements regarding broad data sharing. In doing so, sufficient consideration may not have been

[1]S. J. Quinney College of Law and Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [2]School of Law, Duke University, Durham, NC, USA. *E-mail: jorge.contreras@law.utah.edu

given to potentially useful distributed data structures. When, at the conclusion of a lengthy planning stage, it became apparent that a centralized commons was beyond existing budgetary constraints, the only practical option remaining was to settle for no commons at all and rely on the project's lofty but nonspecific data-sharing principles to motivate researchers to share data on their own (14). To help planners avoid such dilemmas in the future, we offer the following actionable framework for evaluating distributed data commons early in the project-planning phase.

How many data repositories are under consideration? If the number of data repositories is small, then fully distributed, unlinked repositories (i.e., no commons) may suffice. Researchers may easily access each repository, and the cost of implementing a commons structure can be avoided.

Are there resources to develop a common data portal? As the number of data repositories increases, some form of commons structure will likely facilitate data sharing and usage. Although the cost is not trivial, a common data portal can enhance the value and usability of the data. If funding for a data portal is not available, planners may wish to consider a fully distributed commons with legal interoperability.

Are data regulated in the relevant jurisdictions? This question is relevant no matter which commons structure is selected. If data are not regulated or subject to human subject, privacy, health, or similar legal regimes, consider releasing data to the public domain or licensing it under a common-use license. If data are regulated in one or more relevant jurisdictions, planners should consider engaging legal experts to develop a common data access and use policy that complies with regulations in each jurisdiction. For example, if data include human genetic information, both genetic nondiscrimination laws and data privacy regulations should be considered. Legal interoperability, and the ability for users to access and use all data on consistent terms via a single authorization, will be achieved only if the most stringent jurisdiction's regulations are observed in each case or are otherwise addressed (13).

Although the Belmont Forum will doubtless produce a wealth of valuable earth science data, initial appreciation of data-sharing options might have facilitated decision-making and planning among its many national participants and resulted in a more robust data-sharing structure. Addressing these design choices early—while acknowledging budgetary, legal, and political constraints—can save planning and implementation costs later.

REFERENCES AND NOTES

1. J. H. Reichman, P. F. Uhlir, *Law Contemp. Probl.* **66**, 315 (2003).
2. E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge Univ. Press, Cambridge, 1990).
3. B. M. Frischmann, M. J. Madison, K. J. Strandburg, *Governing Knowledge Commons* (Oxford Univ. Press, New York, 2014), chap. 1.
4. Policy RECommendations for Open Access to Research Data in Europe (RECODE), http://recodeproject.eu.
5. International Council for Science, World Data System Strategic Plans 2014–2018, (ICSU, Paris); http://www.icsu.org/about-icsu/strategic-priorities.
6. B. M. Knoppers *et al*., *Genome Med* **3**, 46 (2011).
7. Institute of Medicine, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk* (National Academies Press, Washington, DC, 2015), chap. 6.
8. J. L. Contreras, *Science* **329**, 393 (2010).
9. J. H. Reichman, P. F. Uhlir, T. Dedeurwaerdere, *Governing Digitally Integrated Genetic Resources, Data, and Literature: Global Intellectual Property Strategies for a Redesigned Microbial Research Commons* (Cambridge Univ. Press, New York, forthcoming 2016), chap. 8.
10. J. Palfrey, U. Gasser, *Interop: The Promise and Perils of Highly Interconnected Systems* (Basic Books, New York, 2012).
11. Group on Earth Observations (GEOSS), "Mechanisms to share data as part of the GEOSS data-core" (White paper, GEOSS, Geneva, 2015).
12. P. Murray-Rust, C. Neylon, R. Pollock, J. Wilbanks, "Panton Principles: Principles for open data in science" (19 February 2010); http://pantonprinciples.org.
13. J. Wilbanks, in *Privacy, Big Data, and the Pubic Good*, (Cambridge Univ. Press, Cambridge, 2014), chap. 11.
14. Belmont Forum Steering Committee, "A place to stand: e-Infrastructures and data management for global change research" (30 June 2015); http://belmontforum.org/belmont-forum-governance.

ACKNOWLEDGMENTS

J.H.R. has received support from the National Human Genome Research Institute, NIH, under award no. P50HG003391. J.L.C. and J.H.R. served as members of the U.S. delegation to the Belmont Forum organized by the National Science Foundation. The authors thank anonymous reviewers for constructive suggestions and comments.

		Centralized	Intermediate Distributed	Fully Distributed	Non-Commons
Incremental Research Benefits	**Data access**	Access to all data in unified manner	Access to multiple repositories through central portal	Access to each repository separately, but under a common usage/access policy and single approval	Ad hoc coordination with other repositories only
	Data analytics	Most powerful search, analysis, quality assurance of aggregated data	Cross-repository searching and analytics; Metadata and aggregate statistics can be developed by central authority	Index/Catalog only	Index/Catalog only
Costs	**Up-front costs**	Structure and build centralized repository; Develop data interoperability mechanisms; Develop common usage policy	Develop data interoperability mechanisms; Develop common usage policy	Develop common usage policy	Few up-front costs
	Ongoing centralized costs	Operating and maintaining central repository; administering policies	Operating and maintaining portal; administering policies	Administering policies	Few central costs
	Ongoing distributed costs	Few distributed costs	Operating and maintaining repositories	Operating and maintaining repositories	Operating and maintaining repositories
	Governance overhead	Central repository	Central portal/services, each distributed repository and inter-relationships	Each distributed repository and inter-relationships	Each distributed repository with minimal coordination