INFORMATION SCIENCE

Standards and Infrastructure for Innovation Data Exchange

Data on the global R&D enterprise are inconsistently structured and shared, which hinders understanding and policy.

Laurel L. Haak,[1]* David Baker,[2] Donna K. Ginther,[3] Gregg J. Gordon,[4] Matthew A. Probus,[5] Nirmala Kannankutty,[6] Bruce A. Weinberg[7,8]

Economic growth relies in part on efficient advancement and application of research and development (R&D) knowledge. This requires access to data about science—in particular, R&D inputs and outputs such as grants, patents, publications, and data sets, to support an understanding of how R&D information is produced and what affects its availability. But there is a cacophony of R&D-related data across countries, disciplines, data providers, and sectors. Burdened with data that are inconsistently specified, researchers and policy-makers have few incentives or mechanisms to share or interlink cleaned data sets. Access to these data is limited by a patchwork of laws, regulations, and practices that are unevenly applied and interpreted (*1*). A Web-based infrastructure for data sharing and analysis could help. Data exchange standards are a first step. We describe administrative and technical demands and opportunities to meet them.

CREDIT: MOHSIN RAZA/REUTERS

A Distributed Data Infrastructure

There is no single database solution. Data sets are too large, confidentiality issues will limit access, and parties with proprietary components are unlikely to participate in a single-provider solution. Security and licensing require flexible access. Users must be able to attach and integrate new information.

Unified standards for exchanging data could enable a Web-based distributed network, combining local and cloud storage and providing public-access data and tools, private workspace "sandboxes," and versions of data to support parallel analysis. This infrastructure will likely concentrate existing resources, attract new ones, and maximize benefits from coordination and interoperability while minimizing resource drain and top-down control.

No single organization can or should manage the infrastructure alone. Governments, nonprofits, and for-profits must collaborate, balancing governance and management, if this vision is to be realized in a way that is sustainable and accountable. This balance would need to minimize user costs while enabling all parties to participate on equal footing. There are successful models in the public and private sectors for aspects of this infrastructure, such as the Common European Research Information Format (CERIF) data-exchange standard applied across several European Union countries. These models need to be expanded to be international and interdisciplinary, which has implications for governance and funding. Their cost-effectiveness needs to be rigorously evaluated.

Major data providers, including federal statistical agencies, standards organizations, and private vendors, as well as user communities, should establish a steering committee. The U.S. National Academies Board on Research Data and Information, and the Committee on Data for Science and Technology of the International Council for Science are among the natural sponsors.

Achieving Broad-Based Participation

The multiplicity of players with different objectives—e.g., multinational corporations, nonprofits, government agencies, academic researchers—constitutes a challenge, but they have potentially complementary roles. What is lacking is coordination in establishing and adopting standards and a platform in which data can be combined.

Government funding agencies provide financial support, can encourage development and implementation of coordinated standards in grant and reporting systems, and can require that data produced with their support mesh with the infrastructure where possible. For example, the Brazilian Lattes Platform provides an integrated system to manage research information. Lattes is partnering with CASRAI (Consortia Advancing Standards in Research Administration Information), VIVO (an open research–networking community group), and EuroCRIS (Current Research Information Systems) to identify a shared exchange standard. Coordination of funding and standards could drive analysis standards, upon which additional infrastructure can build.

Corporations understand the value of data sharing and have been creating and curating data that support research on research. These include Web of Science citation data, KNODE's work to link papers and patents, and vendors working with universities to collect and manage research data.

Achieving participation requires an understanding of incentives for users. In the United States, the STAR METRICS initiative could transform research if it is part of a wider set of initiatives, but research organizations are concerned about sharing financial information.

Researchers will need to follow standards for reporting activities, but will benefit from improved attribution of their work and interoperability across reporting systems. Citation standards across disciplines and data types—from data sets to algorithms to organisms—need to be defined and implemented. The Institute for Quantitative Social Science has created citation standards (*2*) that can be used as a starting point for discussions (for example, International Organization for Standardization).

Based on these citation standards, metrics

[1]ORCID, Bethesda, MD 20817, USA. [2]CASRAI, Ottawa, Ontario K1R 7X6, Canada. [3]University of Kansas, Lawrence, KS 66045, USA. [4]SSRN, Rochester, NY 14618, USA. [5]Thomson Reuters Intellectual Property and Science, Rockville, MD 20850, USA. [6]NCSES, National Science Foundation, Arlington, VA 20850, USA. [7]Ohio State University, Columbus, OH 43210, USA. [8]National Bureau of Economic Research, Cambridge, MA 02138, USA.

*Author for correspondence. E-mail: l.haak@orcid.org

Published by AAAS

should be developed that cover nonpublication research outputs. These could be used in a system to incentivize researchers to participate more broadly in data-sharing activities and allow institutions to track the use and impact of the full array of research outputs. These systems must be carefully considered and tested. The goal is to combine research activity and output data from many sources to support the study and understanding of R&D knowledge flow.

The Role of Open Data Standards

The nonprofit sector is well positioned for defining and maintaining data interoperability standards, because it can bring the many data infrastructure players together with minimal conflicts of interest. Several exchange standards are in use. EuroCRIS has been advancing CERIF, which has been adopted by many government funders and research organizations in Europe. In the United States, VIVO has applied semantic Web storage and retrieval methods to research data. Although national efforts show promise, it is critical to create multidiscipline, multijurisdiction, and technology-neutral standards and vocabularies. Open Researcher and Contributor ID (ORCID) provides a persistent registry for uniquely identifying researchers and is working to automate linkages with research activities. CASRAI provides a peer-reviewed, open dictionary of terminology for the semantics and record-structures of research information.

With these underlying exchange standards, the first step is to create a Web-based registry for data, or expand an existing one such as DataCite, to meet the needs of a global, multidisciplinary, multistakeholder community. Specific Web-based user interfaces ("wrappers") can interact with the registry and fulfill discipline-specific requirements, such as metadata and related code needs. In addition to biological and physical sciences research activity data, it is important to include organizations like the Social Science Research Network (SSRN), that focus on social sciences and humanities research to ensure the broadest applicability of the infrastructure, reduce future integration efforts, allow for cross-disciplinary data to be combined in innovative research projects, and assist in identifying interdisciplinary work.

Standard technologies could help achieve initial goals. A wrapper protocol like SWORD (Simple Web-Service Offering Repository Deposit) allows users to communicate with a server. SWORD has been successful for publication repositories, and work to extend it to data repositories should be supported. A common set of XML DTD (eXtensible Markup Language Document Type Definition) tags is needed for consistency and efficiency in content exchanges, simplifying access to data and code stored in various repositories. It is important to provide flexibility, as new technologies will continue to be developed.

A Model for Secure Data Access

Data in this infrastructure should be distinguished in terms of level of sensitivity (*3*). Nonsensitive data (e.g., aggregated or already in the public domain) can be made publicly available. Privacy laws vary between countries; person-level data has the potential to become sensitive if linked or otherwise enhanced. Processes for managing such data need to be implemented. Protocols for each U.S. project, including any person-level data to be deposited, should receive standard Institutional Review Board approval. Sensitive, individually identifiable and institutional data would be housed in a protected enclave. Our preferred model is a Web-based infrastructure to host public and restricted data, in which sensitive data would be restricted to users who had applied for and obtained access.

A public-private working group should develop recommendations for reconciling existing data privacy laws [e.g., (*1*)], state and local laws, copyright law, and international standards. This group should discuss laws and regulations needed for providing access while maintaining the security of administrative data gathered by federal agencies. Such a group could be convened by the Confidentiality and Data Access Committee of the U.S. Federal Committee on Statistical Methodology.

The new NSF National Center for Science and Engineering Statistics (NCSES) secure data access facility provides a model of managed access to restricted data, balancing security and access for cleared researchers. Such models can work for commercial providers, such as the tiered security MarketScan repository, which contains medical claims data and is used to support analysis of health-care cost and treatment and patient behavior.

At least three tools would enhance the infrastructure. First, an automated honest broker approach (*4*, *5*) would allow researchers to request access to data and to perform integrated analyses on multiple data sets housed by different providers. Second, review of findings derived from de-identified data sets to verify and preserve confidentiality is a time-consuming manual process; automation of this statistical analysis would facilitate the research process. Third, statistical code that harmonizes variable definitions across data sets could be shared among users who have been granted access to secure data. These tools will take years to develop; investments in development will prove valuable.

Transforming Research

Researchers lament the lack of data sharing (*6*). By linking data and algorithms to the infrastructure, researchers could—with permissions—access other research projects, encouraging replication and resource utilization. Users would register for access to security-sensitive parts of the infrastructure, use public data and tools free of charge, and pay for access to private work spaces, intellectual property–controlled data sets, or customized analytic tools. Users could post comments on components. Providers would document their data with standard metadata, including data elements, sample frame, access levels, terms of use, and any fees, which could vary according to the amount and nature of use (e.g., scholarly, commercial, or algorithm development), and vary across providers (e.g., academics and government agencies might set prices to cover expenses, with commercial providers setting higher prices for different functionality and tools). Payment could be managed much the way e-readers manage access to applications and content. Such a structure minimizes centralized support and subsidies; allows data to be maintained by providers who can manage access, data updates, and algorithms for data processing; and users can distribute their own tools and algorithms. These objectives are in line with the U.S. government memorandum on data sharing and privacy (*7*).

The proposed model offers potential benefits from combining and mining the vast data already available. The first step is to coordinate existing data exchange efforts, the foundation on which the entire effort relies.

References and Notes

1. These laws include the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), the Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability and Accountability Act (HIPAA), and the European Union Data Protection Directive 95/46/EC.
2. The DataVerse Network Project citation standard; http://projects.iq.harvard.edu/thedata/citation/standard.
3. National Research Council, *Expanding Access to Research Data: Reconciling Risks and Opportunities* (National Academy Press, Washington, DC, 2005).
4. R. Dhir *et al.*, *Cancer* **113**, 1705 (2008).
5. A. D. Boyd *et al.*, *Int. J. Med. Inform.* **76**, 407 (2007).
6. J. P. Mesirov, *Science* **327**, 415 (2010).
7. Office of Management and Budget Memorandum M-11-02, Sharing Data While Protecting Privacy, 3 November 2010; www.whitehouse.gov/sites/default/files/omb/memoranda/2011/m11-02.pdf.

Acknowledgments: We thank J. Saul and J. Hammond for helpful discussion. D.K.G. acknowledges support from NIH grant 1R01AG036820. B.A.W. acknowledges support from NSF grant 1064220. The views and opinions of the authors do not state or reflect those of the NSF.

10.1126/science.1221840