It’s 10 PM; Do You Know Where Your Data Are?
Data Provenance, Curation, and Storage

Mark E. Anderson, Stuart C. Ray

High integrity data retention and curation are critical for preserving the scientific record and informing future discovery.1 However, these steps are often neglected or inadequate because of lack of a tractable, easily operated approach. We offer general guidelines and an exemplar method that is applicable to many, but by no means all, laboratories.

Data Retention and Provenance
Data generated from National Institutes of Health funding should be stored for 3 years after the end of the last competitive renewal. In some cases, data related to patients and patents has longer storage obligations. Data storage rules are in evolution and may differ among various funding agencies, institutions, and journals. The data belong to the host institution, but the responsibility for storage (ie, stewardship) is typically transferred to individual investigators, many of whom have insufficient understanding of or infrastructure for this important role.2 While authors are routinely asked to affirm their accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved,3 there is no standard infrastructure for sharing all primary data among all authors concurrently during drafting of an article, much less after publication.

If you have worked in a laboratory for much time, you will know it is sometimes difficult to locate original data. Some of these challenges are magnified by moving laboratory locations, storing data on proprietary and outmoded platforms, lack of a clear paper trail after laboratory personnel move on, or multiple collaborators generating data at various sites. In general, older data are harder to follow compared with newer data. These truths became starkly evident to me (M.E. Anderson) after a former laboratory member was discovered to have engaged in scientific misconduct. While there was no doubt about his transgressions involving manipulated and repurposed example tracings, published work by us and others, independent of this individual, indicated that his represented findings most likely reflected actual biology. However, the trail of data was incomplete, and most of the publications were over 10 years old. The laboratory notebooks were in hand, but his computer, left behind in a laboratory move, was lost. During the investigation, we repeated many of the key experiments and obtained results similar to those published, but these were unsuitable for replacing the vitiated data because of modern concepts of peer review. These wrenching events led me (M.E. Anderson) and our laboratory to consider improved ways of retaining data, with a focus on a method that was durable and where retrieval of original data used to build published tables and figures did not impose an undue burden and was reliable. In discussions with many scientific leaders, I (M.E. Anderson) learned that few laboratories had clear approaches to storing data, much less ones that would allow data sets immediately supporting publications to be readily located and safely protected. Our laboratory implemented the following simple policy on data storage, sharing, and curation.

Anderson Laboratory Data Sharing and Storage Policy
All data and laboratory records need to be accessible for all laboratory members. Records should be detailed enough for others to recreate an experiment.

Each laboratory member is required to keep the following:
1. A bound notebook describing daily laboratory activities and referencing associated electronic files for each experiment. Records should be written with indelible ink, and errors marked, but not erased. An equivalent electronic notebook is also acceptable.
2. All raw data acquired, either physical (ie, printouts, films, etc) or electronic (ie, images, patch clamp recordings, etc).
3. Backups of all electronic data files. All data should be stored on the laboratory’s shared server drive or backed up on an equivalent secure server (ie, cold storage).

For each published article, the lead author is required to provide the following:
1. The final accepted article file(s).
2. All figures (regular and supplementary) published.
3. All raw data files and calculated data files associated with each figure. The data should be organized into folders by figure.
4. A Word document (or similar) describing the files associated with each figure, indicating any special software.

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD (M.E.A., S.C.R.); and Department of Physiology and the Program in Cellular and Molecular Medicine, The Johns Hopkins School of Medicine, Baltimore, MD (M.E.A.).

Correspondence to Mark E. Anderson, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD. E-mail mark.anderson@jhmi.edu; or Stuart C. Ray, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD. E-mail ray@jhmi.edu

(Circ Res. 2017;120:1551-1554. DOI: 10.1161/CIRCRESAHA.116.310424.) © 2017 American Heart Association, Inc.

Circulation Research is available at http://circres.ahajournals.org
DOI: 10.1161/CIRCRESAHA.116.310424

1551
necessary for viewing files and any additional data that are not electronic (ie, notebooks, films, etc). Other laboratory members should be able to identify the data used to create each figure.

These files should be saved in a folder with the author name, journal, and year published. The folder is to be stored on the laboratory shared server drive.

On leaving the laboratory, all notebooks, data backups, and raw data are to remain in the laboratory with the laboratory manager.

This policy, once adopted, changed practice but did not substantially increase workload because these data had been recently aggregated to prepare the article for publication. At the time data are acquired, proper storage is a matter of practice—like putting away pipettes or wiping down a biosafety cabinet. Data stewardship must be taught and reinforced, but the burden of this practice is reduced if proper IT infrastructure and support is available.

**Different Situations Require Different Solutions**

These guidelines were relatively simple, easy to implement, and provided an approach for data storage in a format purpose for retrieving published data. When a network-attached server (NAS) used to store the data has backups that are retained for years, particularly if logging is enabled, it is feasible to recover the sequence of data changes (additions/deletions/edits) that may have occurred. Given the plummeting cost of deep or cold storage, it seems likely that data so stored will outlast its authors. However, our method did not aim to save for all time the universe of data generated in all studies, did not provide a clear framework for protecting data obtained from humans, nor consider the cost of storing large data sets, such as those generated by imaging studies. After I (M.E. Anderson) moved to Johns Hopkins, I began meeting with data experts to learn more about what our institution was planning and how these plans might align with resources and best practices to serve our faculty and the larger scientific community.

**Recruiting a Departmental Leader**

Our department is research intensive, so approaches to data storage and curation were of broad interest, but a key opinion leader was needed to craft departmental language and speak to faculty at multiple levels. Because institutional policies were changing rapidly to keep in pace with an increasingly complex data security landscape, these changes needed to be communicated to and influenced by our department. Because of this, we established a new Vice Chair for Data Integrity and Analytics and recruited S.C. Ray to lead these efforts and to represent our interests to the School of Medicine and the University. Alignment of departmental efforts with those of the institution is facilitated by participation at the university level in governance of institutional data and clinical applications/analytics.

**What Is Good Enough?**

**Minimum**

- Compliance with sponsor (eg, National Institutes of Health) requirements for laboratory documentation
- Authorship is clear (who wrote what)
- Timing is clear (when they wrote it)
- Completeness (no complete deletion, only strikethrough or equivalent)
- Data/records must remain on campus (removal only with institutional permission)
- If personal health information are included, then data privacy and security rules must be followed
- Only members of the Institutional Review Board–approved study team may have access to personal health information
- Protections against data loss—backups, etc

**Working With the Faculty**

Education and communication are essential. Individuals learn painfully that hard drives fail; this knowledge must be disseminated so that hard drive failure is anticipated. Similarly, electronic devices get lost and stolen; ransomware attacks are now commonplace. Use of commercial storage of sensitive/precious data, without an institutional contract, is not permissible—contracts provide for protection from data loss, breach, and secondary use of data. Most importantly, shared awareness of institutional resources will accelerate adoption of best practices. Threats to data security are constantly evolving, requiring changes that can be frustrating if not clearly justified. Communication between faculty and IT leadership must be bidirectional and collegial; addressing faculty needs and opportunities with respect to data storage can advance science and improve productivity.

**What Assets Are Available to Help Investigators Comply?**

Currently, data storage approaches are highly heterogeneous and many are poorly aligned with risk and regulation (Figure). NAS is readily available, flexible, and powerful. Most institutions secure their networks with lightweight directory access protocols that maintain a database of user credentials, enabling authentication (and loss of credentials on employment termination, etc). Federated authentication or sponsored accounts can give trusted collaborators at other institutions access to shared data. Also prevalent are tools like Active Directory that support creation of groups of users, thereby supporting authorization (eg, study team membership) that can manage shared access. Virtual environments can be augmented to support analytics, at the same time making NAS storage more convenient (by mapping NAS drives to virtual desktops and sharing NAS drives among study team members); in this way, data and tools can be combined without leaving the data center, while still be available to investigators anywhere in the world that has an Internet connection.

**Moving Beyond Minimum Requirements to Best Practices**

A best-practice solution will maximize data integrity at each step in data generation. It is difficult and error-prone to recreate data records post hoc. Moving data to secure storage as early as possible (ideally on acquisition) reduces risk of loss because of storage failure or omission and minimizes additional effort when it becomes routine practice.
What Will Happen and What Won’t Happen?
We anticipate an institutional system that will support investigator-initiated creation of publication/proposal-specific folders similar to that described above for the Anderson Laboratory. These folders would be accessible to all study team members during the creation of figures and draft documents; once published, these resources could be made immutable so that additional files could be added but none removed. The ideal system, like a robust database, would keep a traceable record of all changes—an immutable record providing data provenance. Such a record would be maintained by the corresponding author or their designee and remain available to collaborators in accordance with data retention guidelines.

In a best practice scenario, the primary data associated with a scientific work product will be retrievable, and the contributions of the study team members would be discoverable in the event of an inquiry. The work will be more reproducible because key data and methods will be included. Currently, these elements can be hard to locate with confidence, and such gaps create challenges for future scientific inquiry and for investigating potential scientific misconduct. In our opinion, optimized data storage creates an important obstacle against scientific misconduct, but will not prevent it. Importantly, future discoveries may shed new light on older work, providing an opportunity for additional discovery when the data can be assessed accurately.

Not addressed above are topics that require more specialized solutions. Massive data storage and high-performance computing require specialized hardware, domain-specific design, and decisions with regard to primary and intermediate data that might be retained. Data sharing is an important component of validation and reproducibility of research, with its own set of considerations. Multiinstitution collaborations can be supported by the solutions described above, but may have special requirements. A substantial challenge in maintenance of reproducibility is software/version dependency and proprietary data formats (binary, especially)—considerations that can

Figure. Comparison chart of user features and solutions for storing biomedical research data, based on generalities; specific implementations and institutions will vary, but suggested preference is indicated by color ranging from green (prefer) to red (avoid). aLaptops, if institutionally managed/secured, may be suitable for carrying a copy of research data, but the primary/essential data should reside in a more secure location. bRemote desktop access tools can create vulnerabilities unless managed by IT personnel. cIf primary/essential data are stored, removal from campus would violate policy on location of institutional data (must remain on-site). dRemote access to cloud data requires network access or synchronization (when network access is intermittent); latter may be disabled at some institutions for IT security reasons. eRemote access to network-attached server (NAS) requires network connection and cloud desktop, VPN with drive mapping, or other solution (now becoming routine). fHard drives have high failure rate; laptops increase failure rate and add risk of device loss/theft/damage. gProper authentication and authorization generally depend on device management by institutional IT team. hPrimary/essential research data must remain on campus (or in institutionally contracted storage) and accessible to authorized institutional authorities for compliance reasons. While laboratory data often do not include personal health information (PHI), many laboratories handle clinical specimens with some PHI (medical record numbers, dates of clinical events, date of birth, etc); in addition, clinician scientists tend to have some PHI wherever they work. Under special circumstances when there is not suitable alternative, desktop (locked office server) storage of PHI-containing data may be approved by some institutions. iAffordability of NAS depends on data footprint, economies of scale, institutional investment, Moore’s Law, planning. In the event of concern raised (e.g., to the National Institutes of Health [NIH] Office of Research Integrity), it is essential that the host institution have access to primary data and methods to support the publication or grant proposal in question.
be mitigated by detailed methods (including software versions), scripted analysis, and colocation of software when feasible.

Sidebar
NAS is file-level storage attached to a computing network, routinely available in research settings. With support from network experts, NAS can offer scalable storage with user-level authentication (who is logging in, and are they using valid credentials), authorization (is this user authorized to access this file/folder), event logging (who did what, when), and automated backup.

Acknowledgments
We thank Dr Elizabeth Luczak for her help in developing and implementing the Anderson laboratory data storage policy.

Sources of Funding
This work was supported in part by National Institutes of Health (NIH) awards U19 A1088791, HL079031, HL096652, HL070250, and HL071140.

Disclosures
None.

References
3. ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals 2015.
6. NIH. NIH data sharing policy. 2007.
It's 10 pm; Do You Know Where Your Data Are?: Data Provenance, Curation, and Storage

Mark E. Anderson and Stuart C. Ray

_Circ Res._ 2017;120:1551-1554
doi: 10.1161/CIRCRESAHA.116.310424

_Circulation Research_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2017 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the World Wide Web at:

http://circres.ahajournals.org/content/120/10/1551

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in _Circulation Research_ can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to _Circulation Research_ is online at:
http://circres.ahajournals.org//subscriptions/