VOLUME TWO

CURATING RESEARCH DATA

A Handbook of Current Practice

With 30 case studies contributed by practitioners in the field

BY LISA R. JOHNSTON

Curating Research Data

Volume Two: A Handbook of Current Practice

by Lisa R. Johnston

with 30 case studies contributed
by practitioners in the field

Association of College and Research Libraries
A division of the American Library Association
Chicago, Illinois 2017

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences–Permanence of Paper for Printed Library Materials, ANSI Z39.48-1992. ∞

Cataloging-in-Publication data is on file with the Library of Congress

Printed in the United States of America.

21 20 19 18 17 5 4 3 2 1

Cover image
Copyright: kentoh / 123RF Stock Photo (http://www.123rf.com/profile_kentoh)

- Impactstory (https://impactstory.org) is a tool aimed at the individual researcher (~$60/year) and includes statistics for data from sources such as such as Figshare.com and Github.com.
- Mendeley Data (https://data.mendeley.com) and figShare (https://figshare.com) are data repositories that do a good job at tracking data set downloads and views in ways that increase the value-proposition for researchers to share their data.
- Also of note are ResearchGate, Google Scholar, and Microsoft Academic Search, which are used for tracking data publications (e.g., *Scientific Data*), but their focus is primarily on research articles.

Requests for Access. For restricted-use data, the number of times that access to data is requested may also indicate impact. Placing data behind access barriers is sometimes necessary (e.g., legal restrictions on the data content to project privacy). However, it is still unclear if access restrictions may deter data reuse. Repositories that do require access credentials may be in a better position to follow up with the user to determine if they successfully reused the data. The next case study by Arun Mathur, Johanna Davidson Bleckman, and Jared Lyle illustrates how one repository is helping its reusers make good citation choices in order to best measure long-term impact of the data.

Reuse of Restricted-Use Research Data

Arun Mathur, Johanna Davidson Bleckman, and Jared Lyle[*]

As described in the *Step 1* case study on page 24, the Inter-university Consortium for Political and Social Research (ICPSR) makes restricted-use data available through three highly secure, highly controlled data access mechanisms: ICPSR's virtual data enclave, ICPSR's physical enclave, or using the researcher's approved computing environment with appropriate precautions taken.

Interested parties may apply for restricted-use data electronically via an online request system, which enables ICPSR user support staff to manage individual collection's data use agreements with users.[15] This process includes verifying initial requests (including ensuring system security), transmitting data, tracking data use, and terminating access.

[*] This study is licensed under a Creative Commons Attribution 4.0 License, CC BY (https://creativecommons.org/licenses/by/4.0/).

VERIFYING INITIAL REQUESTS

After researchers apply for data access through an online application system, ICPSR staff review the application submission. This includes ensuring that

- The credentials of applicants match their online identities in institutional directories
- All sections of the application have been completed and the application has been signed by an authorized institutional representative
- For studies that require it, IRB review documentation for the project has been submitted
- The data is a good fit for their research plan; for example, the study is plausible

ENSURING SYSTEM SECURITY

One additional, and important, point of verification for initial access requests involves reviewing the users' security plans for accessing the data. ICPSR provides security plan templates, through which applicants attest that

- Work with the data can be completed only in a secure office by authorized users
- Users may not discuss the restricted-use data in nonsecure or public locations
- Under no circumstances can any unauthorized person be allowed to access or view the restricted-use data, including through windows or doors
- The computer on which the data is viewed must be password-protected and locked if the user leaves, even momentarily
- The computer on which the data is viewed is physically disconnected from the Internet and protected against malware
- Restricted data cannot be copied or duplicated; this includes not taking screenshots or handwritten notes
- If required, users of the virtual and physical enclaves will submit all statistical outputs and results from the restricted-use data to ICPSR for a disclosure review prior to sharing outputs with unauthorized persons
- Users may disseminate only aggregated (i.e., nonconfidential) information from the restricted-use data to anyone not named in the research plan

TRANSMITTING DATA

Once the application is reviewed and all of the criteria are met, the data is then distributed to the user through one of the restricted-use data access mechanisms: via the virtual data enclave (VDE), via the physical enclave, or through a one-time secure download to a researcher's secure environment.

VIRTUAL DATA ENCLAVE (VDE)

A virtual machine is launched from the researcher's local desktop, but the software and data files are operated on ICPSR's server, similar to remotely logging into another physical computer. The virtual machine is isolated from the user's physical desktop computer, restricting the user from downloading files or parts of files to their physical computer. The virtual machine also restricts external access, preventing users from e-mailing, copying, or otherwise moving files outside of the secure environment, either accidentally or intentionally. Available options of the VDE include file sharing among project team members and vetting of output for disclosure risk.

PHYSICAL ENCLAVE

Data is accessible for analysis on-site at the ICPSR building in Ann Arbor, Michigan, in a secure, monitored room. This data is of the highest sensitivity and may contain personal information collected from, for example, victims of violence. When using the enclave, investigators must use non-networked computers provided by ICPSR. The computer cannot send e-mail or access the Internet, and the external media ports (e.g., USB) are disabled. An ICPSR staff member is present at all times when a researcher is using the enclave. The monitor inspects and approves all material brought in or taken out of the enclave, and all output (notes and other material) must be submitted for disclosure review before leaving the physical enclave.

RESEARCHER'S SECURE ENVIRONMENT

Data is provided to the researcher by one-time secure download or by encrypted compact disk. The data must be stored and used only in the computing environment agreed to in the researcher's approved data security plan. The lead researcher is responsible for ensuring that all research team members comply with the security plan and terms of the particular collection's data use agreement. As a security precaution, ICPSR makes each data file unique to the researcher in order to prevent unauthorized dissemination.

TRACKING AND TERMINATING DATA ACCESS

Restricted-use data users accessing data through the VDE or their secure environment are tracked by the online application system. VDE licenses are for one year, with the possibility of renewal. If access is about to expire, the system sends an e-mail notifying users that they may renew access by uploading an annual report and by obtaining IRB approval for an extension, or that they must close out the access agreement.

Upon the end date, VDE remote access is turned off and the user's log-in credentials no longer work. User files are retained for one year to preserve an opportunity for the user to seamlessly renew at a later time. For physical enclave users, work is limited to on-site visits. As mentioned above, usage is closely monitored in person.

Researchers with access via their secure local environment must use a secure erasure program to wipe restricted-use data files and any files containing confidential content from their computer. If the data was transmitted on encrypted compact disk, they must securely destroy the media. Finally, researchers must send ICPSR a signed and notarized affidavit of destruction to attest that all files with confidential data have been destroyed.

SUMMARY

Restricted-use data is available at ICPSR through highly secure, highly controlled data access mechanisms, including a virtual data enclave and a physical enclave. Similar systems are used at other repositories, such as the National Opinion Research Center (NORC) Data Enclave housed at the University of Chicago.[16] Key elements of a request access system include verifying initial requests (including ensuring system security), transmitting data, tracking data users, and verifying end of access.

8.2 Collect Feedback about Data Reuse and Quality Issues

When data is reused, new information is generated about the quality, usefulness, and challenges inherent to the data. Consider incorporating post-ingest review techniques into your repository and curation services that allow others, the general public or subject matter experts, to provide feedback on the data. This process may provide additional post-ingest quality control or aid in the presentation or design of your digital repository.