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)

Making the Case for Disciplinary Data Repositories

*Jared Lyle**

Many advantages come with depositing to a disciplinary repository to make original data available to other potential users. Disciplinary, or domain, repositories "serve a scientific community, which may be a traditional academic discipline, a subdiscipline, or an interdisciplinary network of scientists, united by a common focus."[38] In addition to their core functions of describing, curating, preserving, and providing access to data collections, "they seek to know what [their] community wants and expects in terms of content, format, delivery options, support, and training."[39]

This specialized service is a key differentiator. While institutional and general repositories serve a broad range of users and data, with metadata, access, and user support mechanisms geared to a heterogeneous and wide audience, a disciplinary repository can provide specialized data, services, and tools used and favored by a specific scientific community.

There are several advantages to submitting your data for stewardship in a disciplinary repository.

Subject expertise. "The specialized facility is more likely to know of the existence of important bodies of data relevant to its specialties. Its personnel are best equipped to make judgments as to priorities in data acquisition, as well as to necessary quality controls on new data."[40] Specialized personnel are also best equipped to provide support to reuse the data. The Qualitative Data Repository (QDR, https://qdr.syr.edu), for instance, is a dedicated archive for storing and sharing digital data (and accompanying documentation) generated or collected through qualitative and multi-method research in the social sciences. Trained staff, including international experts in qualitative methodology, administer the repository and understand the unique properties of qualitative and multi-method data.

Customized metadata. "[Disciplinary] archives are familiar with international standards for creating and storing metadata, which greatly enhances the usability, interoperability, and exchange of data."[41] Whereas general self-deposit repositories often solicit minimal metadata, disciplinary repositories can collect and enhance data documentation to ensure complete and self-explanatory collections.[42]

In the social sciences, for instance, many disciplinary repositories use the Data Documentation Initiative (DDI, http://www.ddialliance.org) metadata standard, which captures comprehensive methodological information and enables detailed discovery, including at the variable level. [43]

Disclosure expertise. If data has confidentiality or privacy issues, a disciplinary repository may be "familiar with the [statistical disclosure control] literature and can work with [investigators] to balance the trade-off between data utility and the protection of study participants."[44] For example, the social and behavioral sciences research data repository, the Inter-university Consortium for Political and Social Research (ICPSR, http://www.icpsr.umich.edu), reviews all data for disclosure risk and makes disclosive and sensitive data available through highly secure, highly controlled data access mechanisms, including virtual and physical data enclaves.[45] Likewise, the Digital Archaeological Record (tDAR, http://www.tdar.org), an international digital repository for the digital records of archaeological investigations, reviews and treats confidential and culturally sensitive digital content.[46]

Customized curation and preservation. Disciplinary repository personnel are experts at curating and preserving data, including conducting "data quality reviews" to organize, clean, enhance and preserve data, which "reduces threats to their long-term research value and mitigates the risk of digital obsolescence."[47] Disciplinary repositories are especially attuned to supporting subject-specific formats, which general repositories may not have the expertise or time to handle. For example, Green and Gutmann commented: "The level of long-term support for different kinds of content is an important issue for potential depositors and users of social science data. The …experience of social science data archiving reveals that, while some core formats for datasets have persisted over time, many formats have not."[48] The UK Data Archive (UKDA, http://www.data-archive.ac.uk), the United Kingdom's largest collection of digital research data in the social sciences and humanities, follows a policy of active preservation of acquired content "to ensure the authenticity, reliability, and logical integrity of all resources entrusted to our care while providing usable versions for research, teaching or learning, in perpetuity."[49] In addition to the preservation actions, UK Data Archive acquisitions are curated, which includes data validation, enhanced labelling, grouping of survey variables, and the creation of user guides.

Specialized tools. Because data in disciplinary repositories is curated to a higher level, specialized tools can be offered to further enhance the user experience. At ICPSR, for instance, a Variables Database, containing nearly 4 million variables, enables users to examine and compare variables and questions across studies or series.[50] Users can download data in a wide array of formats (R, SAS, SPSS, Stata, tab-delimited), or they can view and analyze the data on the Web without downloading individual data files.[51] As a final example, ICPSR links each data collection to a list of publications that analyzed the data previously. Citations to these data-related publications are actively collected and managed in a bib-

liographic database.[52] These tools go above and beyond basic functionality found at a general repository.

One-stop, focused collection of data. Disciplinary repositories offer deep collections, making them one-stop sources of data. Similarly, in contrast to general repositories with a diversity of data types, disciplinary repositories "focus on data that benefit from being used in relation to, and in comparison with, other data in the collection."[53] The Archaeology Data Service (ADS, http://archaeologydataservice.ac.uk) in the United Kingdom, for instance, provides a vast catalog of high-quality digital data in archaeology. The Protein Data Bank (PDB, http://www.rcsb.org/pdb) archive is another example of a disciplinary repository as a single source of specialized information—in the PDB's case, 3-D structures of large biological molecules, including proteins and nucleic acids.

SUMMARY

Researchers are fortunate to have so many options for depositing data. Depositing in any trustworthy repository benefits science. That said, disciplinary repositories offer unique and specialized data, services, and tools used and favored by a specific scientific domain and community. These include subject expertise, customized metadata, disclosure expertise, customized curation and preservation, specialized tools, and a one-stop, focused collection of data.

Summary of Step 4.0: Ingest and Store Data in Your Repository

4.1 Ingest the Data Files: Transfer the processed data files to the repository while maintaining integrity and verifying fixity throughout the process (e.g., generate checksums of the files).

4.2 Store the Assets Securely: Add the ingested files to a well-configured (in terms of hardware and software) archival storage environment. Perform routine checks and provide disaster recovery capabilities as needed.

4.3 Develop Trust in Your Repository: Become a trusted digital repository for data by applying for accreditation and growing your reputation locally and beyond.

Notes

1. These functions are described in useful detail from Magenta Book's Recommended Practice manual (Consultative Committee for Space Data Systems, *Reference Model for*