Title
Data Sharing, Storage, and the Social Sciences Working Group

Authors
Wittenberg, Jamie
Church, James
Dekker, Harrison
et al.

Publication Date
2021-12-08

DOI
10.25350/B55P4H

Data Sharing, Storage and the Social Sciences Working Group

Recommendations and Report

December 9th, 2016

Working Group Membership:
Jamie Wittenberg (Chair)
Jim Church
Harrison Dekker
Celia Emmelhainz
Jon Stiles

Table of Contents

Summary of Charge

Social scientists increasingly need to preserve and share their research data in systematic ways. Yet researchers at Berkeley are managing data on many different personal or institutional websites--or not at all--and the Library is not systematically advising on best practices. Given that data sharing and data preservation align closely with the role of the Library in sharing, preserving, and making data : to share, preserve and make accessible.

This working group's charge was to analyze options for social science researchers at Berkeley to preserve and share their data. We have gathered input from researchers and peer institutions, examined the pros and cons of available data archiving options, and made "best practices" recommendations for different types of data.

Top Recommendations

Recommendations for Researchers

In evaluating options for data storage and sharing, we discovered a number of different potential use cases. Our recommendations for researchers are displayed in a flowchart on the next page, as well as in a textual format here:

Publisher Requirements

1) Meet all publisher or funder requirements or guidance first. If they are not specific, see below:

Sensitive Data

2) If your sensitive data is within scope for ICPSR, place in **ICPSR**.
3) If your sensitive data is not within scope for ICPSR, place metadata in Dataverse but deposit the data in a **dark archive**.
4) If your sensitive data can be de-identified, do that and treat as non-sensitive data.

Non-Identifying Data

5) If you don't require variable-level indexing, place in **Dash/Merritt**.
6) If you need variable indexing and want data open-access...
    a) ...and its use value is specific to the social sciences, place in **ICPSR Open**
    b) ...and it has use value outside of the social sciences, place in **Dataverse**
7) If you need variable indexing and do not mind closed-access, place in **ICPSR**

Recommendations at the Institutional Level

At the institutional level, we recommend:

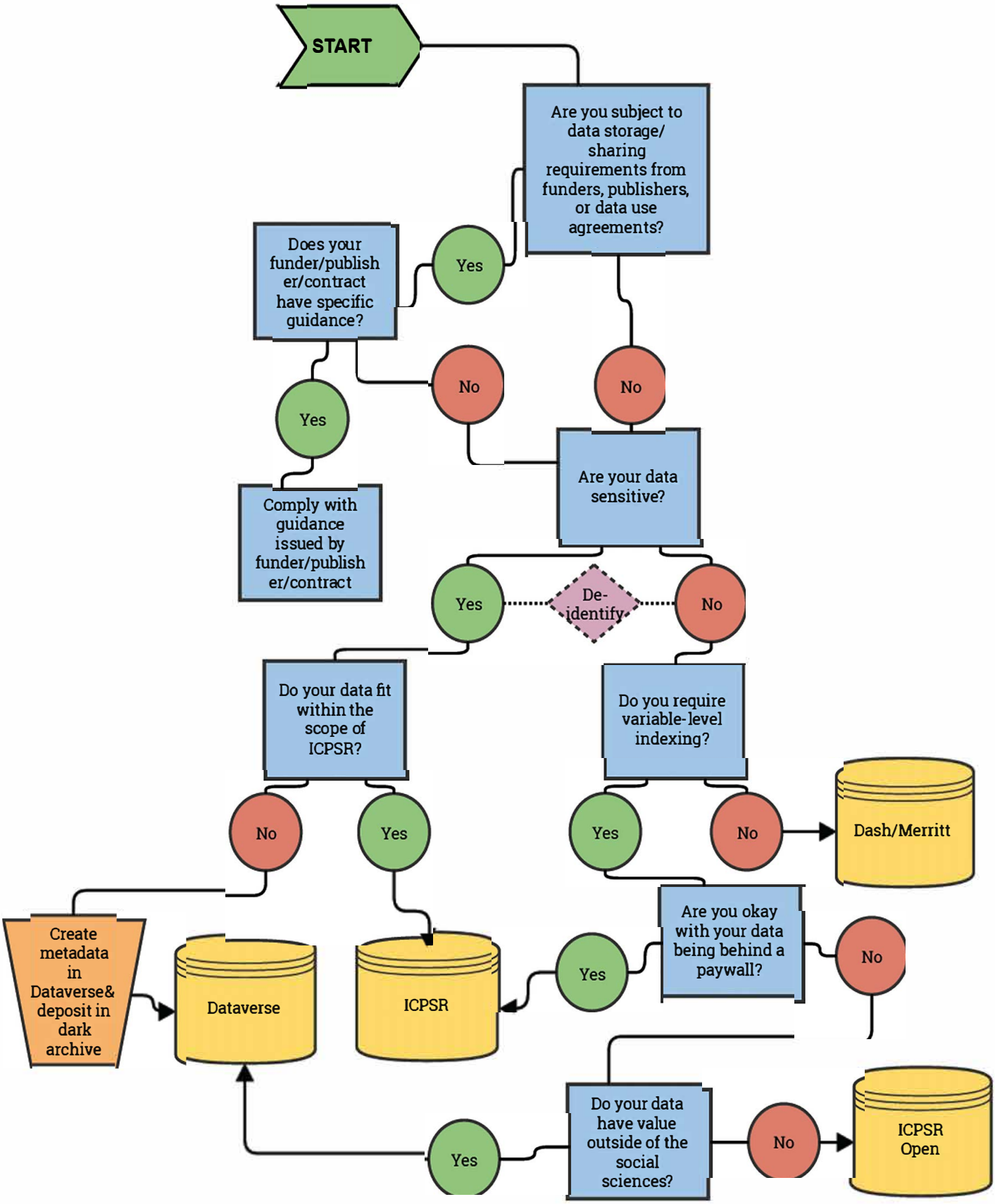
- Reevaluating available resources for data storage and sharing on a yearly basis.
- Working groups in other library divisions to consider options in their domains.

Creating a local Dataverse to archive the following use cases:

- Data that are sensitive and do not fit within the scope of ICPSR
- Data that are not CC0 licensed (e.g. need a more restrictive license for open data)
- Data owners who want or require more granular permissions (ie, embargo, provision on request)

doi:10.25350/B55P4H

Social Science Data Archiving Recommendations Decision Tree



Report on Available Data Sharing Options

ICPSR

For social science research, ICPSR is our recommended repository. Compared to other repositories, it is well established and widely recognized, maintains preservation, accessibility, and metadata standards, and is one of a handful of US repositories that currently holds the Data Seal of Approval.

There are two options for depositing data: assisted deposits which are checked through our institutional membership with ICPSR for both data access and depositing (which currently costs UCB $17,400 a year). Researchers can also deposit into openICPSR, which is not moderated but lets them make data accessible worldwide. A institutionally branded version of openICPSR is available on a fee basis.

For sensitive (e.g. interviews, fieldnotes) qualitative or quantitative research data, ICSPR offers some restricted archiving, although online analysis is not possible and (appropriately) researchers need to sign re-use agreements for secondary research. Access to restricted use quantitative data is provided under several modes, including through ICPSR's Virtual Data Enclave. If ICPSR determines that sensitive research falls out of scope, they refer researchers to the Qualitative Data Repository (QDR) at Syracuse, which is still in beta.*

ICPSR is a founding member of DataPASS, an agreement with large institutional partners to provide collection preservation and backup. ICPSR is a member of the Dataverse consortium, and its holdings are searchable and accessible through Dataverse.

*QDR is still in beta and not recommended for data archiving at this point. There is no major qualitative data archive in the US that meets the Data Seal of Approval; UKDA (the UK Data Archive) has a strong qualitative archiving program in the United Kingdom, and several other European countries also host well-designed qualitative archives.

Figshare

Figshare is a commercial data publishing and discovery platform funded by Digital Science. Digital Science is operated by global media company, the Holtzbrinck Publishing Group. Figshare is particularly effective as a tool for access and facilitates rendering, embedding, citing, and referring to digital objects. Figshare offers a free consumer option and an enterprise option.

At present, there are 60 users with @ucberkeley.edu domain using the Figshare site. Figshare imagines there is probably 3 times this number based on personal email patterns. Of these 60 authors, they have created 351 items, of which 82 are public. They have also created 60

collections and 15 projects, all of which are being worked on privately. Enterprise instances of Figshare sit on top of repositories, and so preservation is dependent upon the underlying repository. The consumer version of Figshare saves all files on Amazon S3 and saves backups that are retained for 5 days, as well as weekly snapshots of the entire data system, including the Amazon S3 file store.

Github

Github is a commercial Git repository hosting service that offers free consumer and subscription-based enterprise accounts. Github is primarily a collaboration and accessibility tool and is not recommended for data storage. However, it is highly effective for data sharing and especially for sharing code. Consumer Github imposes a 100 MB maximum on single files, and recommends repositories be under 1 GB. This makes it potentially unsuitable as a sharing mechanism for social scientists with video data. (Git-Annex does support “arbitrarily large” data files, however.) Github should not be used for medium or long-term archiving as it meets almost none of the Data Seal of Approval criteria.

Dash/Merritt

Dash is a data publishing service that sits on top of Merritt, a preservation repository. Dash/Merritt are developed and managed by the University of California. Dash and Merritt are core service offerings of UC3, which is the CDL program supporting the long-term preservation of, and access to, the University’s digital content. UC3 is dedicated to providing long-term preservation of UC digital assets. Dash relies on Merritt for long-term preservation. Merritt uses geographic replication as a primary strategy for preservation, currently hosting storage in UC private cloud services at SDSC and UCLA. In Nov 2016, these will be augmented by integration with Amazon AWS S3 and Glacier storage.

The Merritt repository runs a continual process of audit verification on all of its content, including Dash datasets, to ensure that all stored replicas are faithful copies of each other. In the event that bit-level damage is identified, the damage is corrected by replacing the damaged copy with another verified copy. Data contributed to Merritt can be accompanied by a producer- verified checksum value. This value is then continually verified through the auditing procedure described in the previous answer. Merritt maintains a complete change history of all data, and any prior state can be re-instantiated and retrieved. There are no prescriptive format requirements; Dash/Merritt can accept any type of data in any format.

The UC3 group are in the process of applying for a Data Seal of Approval. Dash/Merritt are recommended for medium and long-term archiving of de-identified social science data. Because all data deposited with Dash are discoverable, data must be de-identified prior to deposit. Dash does have some features to facilitate sharing, like CC-BY licensing, DOI minting, and indexing. However, it is not recommended for providing access copies of media files like video and images, as it will not render them and will require download.

UC Data Archive & Technical Assistance

UC DATA was the historic social science data archive for UC Berkeley. Started in 1959, it initially specialized in survey data collected in Latin America and Africa, and maintains unique specialty collections of data from the Census Bureau, historic higher education data collection, and California polls. As the representative for ICPSR, UC DATA focuses on and assists in placing researcher data into that supported repository as a first preference, but can locally archive and preserve researcher data on request. In addition to documentation and preservation, UC DATA can also place quantitative data into SDA, an online analysis tool also by IPUMS and ICPSR. UC DATA houses copies of its holdings with the data archive for UCLA as a preservation policy.

UC DATA is currently under-resourced and not recommended as a repository for new researcher-initiated data deposits that can be otherwise deposited in other available full-service repositories, but can provide assistance to researchers who want help with preparing their data for deposit. It is most suitable for deposits oriented toward needs for online analysis and access, or larger thematic collections that can attract external support.

Dataverse

The name Dataverse refers to an open source software project headed by the Institute for Quantitative Social Science and Harvard University. The purpose of the Dataverse software is “to share, preserve, cite, explore, and analyze research data.” Development of Dataverse has been ongoing since 2007 and builds upon an earlier project called Virtual Data Center (1999-2006). The work is funded by Harvard University as well as grant support from such agencies as Alfred P. Sloan Foundation, National Science Foundation, National Institutes of Health and many others.

A Dataverse software installation provides a central repository which can host multiple “dataverses.” A dataverse can contain data and metadata files or can serve as a container for other dataverses. Typical use cases are individual dataverses for sharing a single researchers their data and institutionally managed dataverses for sharing the work of multiple researchers. Quality assurance, adherence to metadata standards, and ingest workflows are the responsibility of the dataverse owner. The software provides flexible support for access restrictions and terms of use agreements, but again, it’s the responsibility of the dataverse creator to make the correct determination of which to use.

Harvard offers free and paid hosting plans on the Harvard Dataverse to both individuals and institutions. A number of institutions run their own instances, including the University of North Carolina at Chapel Hill, University of Virginia, and the University of British Columbia.

Results of Survey

A short anonymous survey was created and fielded to faculty researchers via requests from Library area leads. In total, only 11 respondents replied to initial requests. We did not pursue repeated contacts or strategies to expand our pool of contacts for faculty researchers. Collaborating with the Sponsored Projects Office or the Research Data Management group in Research IT to integrate this survey of researcher practices would likely improve our understanding of faculty needs.

The short questionnaire focused on identifying the collection and use of primary and secondary data by faculty. For researchers who collect their own data, we asked about their data sharing practices, any challenges they face in sharing their data, and venues through which they make data available. For researchers who do use secondary data, we asked about how they heard about those resources and how they obtain those data. For background purposes we also asked about the department and field they are associated with and do research in. We did not force responses to any question.
The Qualtrics questionnaire is included in Appendix C.

Of the 11 respondents, 3 researchers indicated they collect their own primary data, 3 indicated they use secondary sources (in 2 cases exclusively), and 3 do not use data in their research. (Three failed to respond to the question on personal practices). Of the three respondents who collect their own data, one shared data with direct collaborators and one shared data with other colleagues who are researchers. None describe their sharing of data through archives or institutional repository, citing concerns over the time and resources needed to adequately document the data, or desire to personally vet the researcher requesting access. However, both for sharing their data/results and for use of other researchers' data, ICPSR was cited, as was Dataverse, figshare and direct contacts with data collectors.

Given the lack of systematic sampling and low level of response, we view the survey primarily as an initial exercise in developing a better approach to understanding researcher needs in data management and preservation. We see at least four areas for improvement: developing a more extensive and well-defined pool of researchers who gather and use data for their research; leveraging partnerships to increase response rates; clarifying through survey examples the distinctions between archives and other methods of data distributions, and; distributing the survey in a non-anonymized fashion, to allow for non-response followup and evaluation of non-response.

Cost Analysis

Service	Consumer version cost	Enterprise version costs	Cost to Researcher	Notes
Figshare	Free	/year over 2 years	Free	Year 1: $ Year 2: $
Github	Free	$ per 10 users / year	Free	Each additional user costs $ /month.
ICPSR	Free*	$ for 1-49 deposits (up to 50GB), more for higher tiers. Branded.	Free	Includes support for restricted use data, world's largest archive of behavioral and social science research data, DDI metadata. [1]
QDR	Free	No cost; still in beta	Free	Has limited capacity and still in beta
Dash	N/A	$ per tb per year, currently 11 datasets, 15 versions, 735 files, 1.6 GB = /GB/year	Free	Currently $ / year
UC Data Archive	Free**	For SDA access, institutional cost of $	Free	Not currently staffed to support substantial increase in deposits
Dataverse	Free	No cost for software, administrative costs for operating and startup costs for hardware	Free	UC Davis implementation does not require additional staff

[1] "openICPSR is undergoing development and is currently free for all users to share their data up to a 2GB limit"

Challenges Going Forward

Data sharing and curation is a complex set of services that require a variety of skills and knowledge, and is evolving due to technological change, shifts in policies and incentives to share, and changes in the nature of data being shared. Identification of appropriate partners on campus and elsewhere to draw on needed and current skills, while ensuring longer-term sustainability will require cultivation and maintenance of those partnerships. Given the complexity of the landscape, the working group recommends that next steps include:

- Establishing a committee to explore the implementation of these recommendations
- Developing guidelines around the minimum documentation and metadata requirements for social science data repository submission
- Further explore the costs of curation for the Library
- Explore the possibility of leveraging tools and services the Library already employs to manage its own collections
- Explore the new features developed as part of Dash v2

References

Faniel, Ixchel M., Adam Kriesberg, and Elizabeth Yakel. 2016. “Social Scientists’ Satisfaction with Data Reuse.” *Journal of the Association for Information Science and Technology* 67 (6): 1404–16. doi:10.1002/asi.23480.

Green, Ann G. and Myron P. Gutmann. 2007. “Building Partnerships among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives.” *OCLC Systems & Services: International Digital Library Perspectives* 23 (1): 35–53. doi:10.1108/10650750710720757.

“Guidelines 2017-2019, Nov. 10, 2016.” 2016. Accessed November 29. https://assessment.datasealofapproval.org/guidelines_54/html/.

“Implementing a Data Citation Workflow within the State Politics and Policy Quarterly Journal | IASSIST Home.” 2016. Accessed November 29. http://www.iassistdata.org/conferences/2015/presentation/7064.

“Qualitative Data at the ICPSR Social Science Data Archive.” 2016. *Databrarians*. March 28. http://databrarians.org/2016/03/qualitative-data-at-the-icpsr-social-science-data-archive/.

“Show Me the Money – the Path to a Sustainable Research Data Facility | Unlocking Research.” 2016. Accessed November 29. https://unlockingresearch.blog.lib.cam.ac.uk/?p=631

doi:10.25350/B55P4H

Appendix A: Comparison of Storage and Sharing Platforms

Guidelines (Updated 2017-2019)	**Figshare**	**Github**	**ICPSR (Qualitative)**	**Qualitative Data Repository (QDR)**	**Dash/Merrit**	**UC Data Archive**	**Dataverse**
Type	*Commercial*	*Commercial*	*Qualitative*	*Qualitative*	*Institutional*	*Institutional*	*Institutional*
Sources consulted	figshare com & Mark Hanhel		*ICPSR site, plus interview (http://databrarians.org/2016/03/qualitative-data-at-the-icpsr-social-science-data-archive/)*	*https://qdr.syr.edu/*	Perry Willet & Stephen Abrams		dataverse.org
ensures data is created, curated, accessed, and used in compliance with disciplinary and ethical norms	Not really, but they're working on creating disciplinary repos	No	strong understanding of privacy; responsibility for anonmization on producer	yes; mostly poli sci at present	Primary responsibility for direct interactions with data producers, about ethical norms as well as other issues, is held by campus library staff.	Yes, primarily around areas of confidentailty and accessability	Self-curation model, so basically no, but has a number of features that if used will facilitate compliance.
	"figshare accepts any file type"	No	yes, extensive documentation on best practices	yes, growing documentation and recommendations	There are no prescriptive format requirements; Dash/Merritt can accept any type of data in any format. UC3 staff are available for consultation regarding format choices that may be preferred.	Recommended formats, primarily oriented toward traditional (non-video, auidio) data.	Many formats supported. Includes built-in utilities for exploration and manipulation of a variety tabular formats and geospatial formats.
accepts data and metadata based on defined criteria	Yes, but very limited	No	yes	yes	Dash requires four elements of the DataCite metadata schema: creator(s), title, abstract, and format type. All other DataCite elements may be optionally supplied. Non-DataCite metadata may also be supplied as independent files making up the dataset.	Yes	Yes. "Committed to using stard-compliant metadata." three levels of metadata supported: citation, domain-specific, file-level.
explicit mission to provide access to and preserve data	Yes	No	yes	yes, but more in sharing	Dash and Merritt are core service offerings of UC3, which is the CDL program supporting the long-term preservation of, and access to, the University's digital content.	Yes, but limited staffing and capacity	Yes
due diligence to comply with laws and contracts, as well as protection of of human subjects	Doesn't do much here	No	responsibilty on producer	responsibility on producer	All Dash webpages, have a Terms and Use link that points in turn to CDL's Terms of Use and Privacy policies. We discourage submission of datasets with personally identifiable information.	As part of consultation with data producers	No, but depostors can create customs terms of use/access and can restrict access.
applies documented processes and procedures in managing archival storage	Yes, uses Amazon S3	Yes	yes; established data management	not visible on website	The Merritt repository relies on a distributed storage broker architecture. The specifications for this architecture, also describing data workflows, are publicly available.	No	Yes. Stated commitment to best archival practices.
has a plan for long-term digital preservation / NEW: continuity plan to ensure ongoing access to and preservation	Sort of - minimum of 10 years for publishers, as long as repository exists for everything else	No	yes	claims yes; not specified	UC3 is dedicated to providing long-term preservation of UC digital assets. Dash relies on Merritt for long-term preservation. Merritt uses geographic replication as a primary strategy for preservation, currently hosting storage in UC private cloud services at SDSC and UCLA. In Nov 2016, these will be augmented by integration with Amazon AWS S3 and Glacier storage.	Yes, via agreements with other archives	Yes.
Archiving takes place according to defined workflows from ingest to dissemination	Sort of	Yes	yes	still in beta	Dash is intended for self-service operation, so its exact position within external researchers' workflows is not pre-determined.	No recently updated flows specified	No. Self-curation model so depends on the depositor.
"assumes responsibility for long-term preservation and manages this function in a planned and documented way "	Yes	Yes	yes	yes	All data submitted to Dash are indexed for public search, display, and retrieval.	Some, but not all.	Yes. Deposited data is never de-accessioned except under extreme circumstances (e.g. legal)
enables users to discover the data and refer to them in a persistent way through proper citation	Yes, collabs with datacite and mints DOIs	Yes to discover and use, no real support for citation	yes; good search interface	yes; move to membership model this fall	All data submitted to Dash are assigned DOIs for permanent citation. Discovery can happen through Dash's on search interface, via DataCites's search interface, or via internet search engines, which index all Dash content. Some Dash content is also registered with the DataONE network, and can be discovered via DataONE's aggregated search interface.	Yes, for a subset of its holdings. Varies by collection	Yes
	Regular fixity checks	No	yes; runs checksums etc	not visible on website	The Merritt repository runs a continual process of audit verification on all of its content, including Dash datasets, to ensure that all stored replicas are faithful copies of each other. In the event that bit-level damage is identified, the damagey is corrected by replacing the damaged copy with another verified copy.	LOCKSS	Claims support for "permanenet bit--level preservation".
guarantees the integrity and authenticity of the data	No	No	(not sure what this means)	?	Data contributed to Merritt can be accompanied by a producer- verified checksum value. This value is then continually verified through the auditing procedure described in the previous answer. Merritt maintains a complete change history of all data, and any prior state can be reinstantiated and retrieved.	No	No
has expertise to address technical and metadata quality; ensures end users can make quality-related evaluations [no mention of OAIS]	Does not meet all requirements established in OAIS model	No	yes (not confirmed)	not visible on website	The design, implementation, and operation of the Merritt repository is consistent with the OAIS reference model.	No	No
sets clear access regulations for data consumers / NEW: "enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data"	Yes	No	yes, in part subscription model	yes; user controls available	All Dash data are available for public display and retrieval.	For a subset of the collections	Supports long term access to data but no ensurance of appropriate metadata
sets clear codes of conduct for data consumers / NEW: "enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data"	No	No	yes	yes	All Dash datasets are associated with an explicit license defining the terms of their use. Licensing information is prominently display on the dataset landing page. Most Dash content is licensed under the Creative Commons CC-BY license, although some is under CC0, and another small fraction have explicit data user agreements.	For a subset of the collections	Defaults to CC0 license, and supports custom terms of use and restricted access.
maintains all applicable licenses covering data access and use and monitors compliance	Yes	No	yes	yes	See previous answer.	For a subset of the collections	No compliance monitoring.
has adequate funding and sufficient numbers of qualified staff			*note: refers to QDR if not in scope for ICPSR*	*note: refers to UKDA, ICPSR, ADS, etc as models*			Supported by Harvard University and has good track record of procuring funding from NSF, Sloan, NIH, etc.
has mechanism to ensure (in-house or scientific) expert guidance			*note: for both of these, responsibilty of data management, anonymization fully on producer*				
uses supported OS and appropriate infrastructural hardware and software			****ICPSR has the data seal of approval*				
technical infrastructure protects the facility and its data, products, services, and users							

Appendix B: Social Science Journal Data Policies

Economics

The American Economic Association (AEA) requires authors to "publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the Review, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the AEA website." *The Journal of Political Economy*, the *Journal of Labor Economics* and *Quarterly Journal of Economics* have adopted the same standard set by the AEA.

Political Science

The American Political Science Association *Data Access & Research Transparency Policy* "requires authors to ensure that cited data are available at the time of publication through a trusted digital repository. Journals may specify which trusted digital repository shall be used (for example if they have their own dataverse)." The policy notes that if data are restricted authors must notify the editor at the time of submission. In an article *Data Availability in Political Science Journals* published in the journal *European Political Science,* the authors also find that most data political science journal data publication policies are general (e.g. do not specify the moment when data has to be submitted), inclusive (referring both to qualitative and quantitative data), specific in the procedures to be followed, and strongly enforced (presented as mandatory for contributors). It was also found that journals with higher citation counts are more likely to have a data availability policy than publications with fewer citations and lower impact.

Sociology

The American Sociological Association (ASA) code of ethics (2008) asserts that, "Sociologists share data and pertinent documentation as a regular practice… [and] anticipate data sharing as an integral part of a research plan…. whenever data sharing is feasible" and that "sociologists who do not otherwise place data in public archives keep data available and retain documentation… for a reasonable period of time after publication." The aspirational nature of these statements may be the reason the code is not widely observed. In the article *Data Sharing in Sociology Journals (2014)* the web sites of 140 sociology journals were consulted to check for data publishing policies. Only a few sociology journals were found to have explicit data policies, with most journals referring to a common policy supplied by their association. Among the journals surveyed, few articles provide data citations and even fewer make data available, for both for journals with and without a data policy. Authors writing for journals with higher impact factors and with data policies are more likely to cite data and to make it accessible.

Psychology

In 2015 the American Psychological Association (APA) convened a data sharing working group that prepared a report entitled *Data Sharing: Principles and Considerations for Policy Development*. In it they observe:

- Sharing data promotes scientific progress.
- Sharing data within the larger scientific community encourages a culture of openness and accountability.
- Sharing data allows geographically dispersed individuals and those with limited resources to investigate scientific questions of interest.
- Sharing data promotes aggregation for the purposes of knowledge synthesis, hypothesis generation, programmatic decision-making, and generalizability testing.

The APA takes particular caution to protect the rights of human subjects and also states that “research and academic institutions and scientific publishers should establish standards for data management and sharing and for storage and preservation of data in secure repositories”. This is an example of another aspirational “should statement” that places the data sharing responsibility on the shoulders of the institutions and publishers, and is not an actual requirement of the association or its members.

Appendix C: Faculty Survey of Data Management and Storage Practices

Survey Questionnaire

How do you find, collect, use and share the fundamental evidence that underlies your research?

1. Do you collect your own primary data (e.g. from surveys, trials, interviews or experiments) for your research? (either alone or in addition to use of other secondary data).
    ___ Yes, I collect my own primary data (1)
    ___ No, I use only secondary data for my research (2)
    ___ I'm not sure (3)
    ___ I do not use data for my research (4)

(IF they do not use data for research, skip to end of survey)

2. Have you made the primary data you collect available to other researchers? Check ALL that apply.
    ___ Yes, I have directly shared with research collaborators
    ___ Yes, I have directly shared with other colleagues or researchers
    ___ Yes, I indirectly share through an archive, repository or website
    ___ No, I have not shared primary data but plan to in the future
    ___ No

3. What concerns or needs do you have about finding, getting access to, or using research data collected by others?

IF they do not share primary data through an archive, repository, or website:

4. Is there a reason you have not shared primary data through an archive or repository? *(Check ALL that apply)*
    ___ No time or resources for adequately documenting data
    ___ Cost of maintaining external access too high
    ___ Unsure of best place to use for redistribution
    ___ Desire to vet quality of researchers working with my data
    ___ Need to protect confidentiality of respondents/objects of research
    ___ Other reason(s) - please specify: _____

IF they do share primary data through an archive, repository, or website:

5. Through what archives, repositories or sites have you made your research data available? *(Check ALL that apply)*
    ___ ICPSR: The Inter-university Consortium for Political and Social Research
    ___ DASH (UC Institutional Repository)
    ___ Dataverse
    ___ Dryad (8)
    ___ Personal website
    ___ Project website
    ___ Figshare
    ___ Other: _____

6. Do you use data collected or created by other researchers for your own research? (either alone or in addition to use of your own primary data)
    ___ Yes, I use secondary data created or collected by others
    ___ No, I use only primary data for my research
    ___ I'm not sure
    ___ I do not use data for my research

IF they use data collected or created by other researchers:

7. From what sources have you obtained secondary data for your own research? *(Check ALL that apply)*

    ___ Directly from research collaborators
    ___ Directly from other colleagues or researchers
    ___ Directly from data producers or "owners"
    ___ From a data archive or institutional repository (please specify):
    _____
    ___ From some other source(s) (please specify) _____

8. How did you find out about these data?

9. With what department(s) or research unit(s) do you identify

10. In what general research area domain do you do research?
    ___ Health and Life Sciences
    ___ Social Sciences
    ___ Engineering and Physical Sciences
    ___ Arts & Humanities
    ___ Other: _____

11. What concerns or needs do you have about sharing your research data?

12. Would you be willing to talk with us further about your needs and uses of research data? If so, please enter your name and a contact email or phone number in the space below.

Appendix D: Data Seal of Approval Guidelines

Data Seal of Approval: Summary of Updated 2017-2019 Guidelines

The repository:

1) has an explicit mission to provide access to and preserve data in its domain.
2) maintains all applicable licenses covering data access and use and monitors compliance.
3) has a continuity plan to ensure ongoing access to and preservation of its holdings.
4) ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
5) has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
6) adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).
7) guarantees the integrity and authenticity of the data.
8) accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
9) applies documented processes and procedures in managing archival storage of the data.
10) assumes responsibility for long-term preservation and manages this function in a planned and documented way.
11) has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
12) Archiving takes place according to defined workflows from ingest to dissemination.
13) enables users to discover the data and refer to them in a persistent way through proper citation.
14) enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
15) functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
16) The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.