Best Practices for Data Sharing and Deposit for Librarian Authors

Regina Fisher Raboin
University of Massachusetts Medical School

Et al.

Let us know how access to this document benefits you.

Follow this and additional works at: https://escholarship.umassmed.edu/jeslib_presentations

Part of the Scholarly Communication Commons, and the Scholarly Publishing Commons

Recommended Citation

Creative Commons License

This work is licensed under a Creative Commons Attribution 4.0 License. This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in Journal of eScience Librarianship Presentations by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.
Best Practices for Data Sharing and Deposit for Librarian Authors

Thank you to Kate Nyhan and NEASIST, and Cathy Nash and Rodneikka Scott and ASIST for inviting us to present on this important topic.

Now let’s get to the heart of the matter.
Today’s webinar is designed and presented by the Journal of eScience Librarianship’s Editors Julie Goldman, Managing Editor; Lisa Palmer, Distribution Editor; and me, Regina Raboin, Editor; and we are joined by T. Scott Plutchak, Editorial Board member for JeSLIB (Journal of eScience Librarianship) and JMLA (Journal of the Medical Library Association).
Learning Objectives

1. Interpret the data sharing policy of a journal
2. Identify different platforms for sharing data
3. Be aware of best practices for sharing data

The learning objectives for this webinar are designed for librarians as researchers and authors focused on depositing data to accompany articles and other works submitted for publication.
The content we will be covering today is:

- A brief outline of why data sharing is important, incentives and impact of funder policies and journal policies
- Everyone knows what data is, but it's important to understand that the definition of data is changing and overtime journals will evolve their policies, different journals define “data” differently; distinction between supplementary materials and data, and authors need to know which journals to target when submitting manuscripts.
- Review the new JMLA open data policy and creation
- A brief comparison of ‘data deposit requirement policies’ of library science journals to selected science journals
- Best practices and how to approach data sharing
- Explore examples of supplementary materials in JeSLIB and explore specific examples of data deposit in JeSLIB and JMLA
- Questions & Discussion
- And a list of resources will complete the webinar
Overview of Data Sharing

T. Scott Plutchak, MA, AHIP

Editorial Board, JeSLIB & JMLA
University of Alabama at Birmingham (Retired)
Data Sharing

Simple definition:

“The practice of making research data available to others for validation and replication of results.”

-- New England Collaborative Data Management Curriculum

There are both simple and complex definitions to the term “data sharing,” but the bottom line is that the practice of data sharing brings a higher level of integrity to research (via the ability to validate and replicate results), speeds the research process by saving others the time to generate data that’s already available, and allowing funders the ability to access the data behind the research they support.
Rise of Data Sharing

- Under discussion for many years, mostly in the hard sciences
- Increasing interest from funders
- Increasing interest from journals
- Likely that researchers in all disciplines will need to comply with data sharing policies

The past few years have seen an increasing wave of interest in open data and in policies and practices for sharing research data. Although this interest is primarily in the hard sciences, increasing interest is being seen in the social sciences and humanities as well.

Related to “Increasing interest from journals,” the recent Scholarly Kitchen article about Springer Nature and data sharing highlights recent developments: https://scholarlykitchen.sspnet.org/2019/01/30/guest-post-encouraging-data-sharing-a-small-investment-for-large-potential-gain/
• NSF has had a modest data sharing policy for years, although some Directorates adopted early, more prescriptive policies and practices. As the building of subject and agency-specific repositories increased, data sharing policies have evolved. NIH had established a minimal data sharing policy for very large grants and some of the individual programs had specific policies.

• Following the release of the “Holdren memo” (White House Office of Science and Technology Policy) in February of 2013, all US Federal Funding agencies have developed plans for requiring their grantees to develop plans for sharing their research data.

• These have been developing slowly – NIH just recently held a request for comment period for their draft data sharing policy – but the trends are clear.

• Similar policies have been implemented or are being developed at other funding agencies and foundations as well.
In response to these trends and policies, many journals have started to develop policies requiring authors to submit plans for sharing the data underpinning the conclusions of their published articles. Unfortunately, that development has been haphazard, resulting in quite different policies and varying processes for implementation and oversight.

PLOS One: https://journals.plos.org/plosone/
Scientific Data: https://www.nature.com/sdata/
GigaScience: https://academic.oup.com/gigascience
Science: https://www.sciencemag.org/
Springer Nature: https://www.springernature.com/gp
In an attempt to provide some guidance for journals seeking to develop such policies, the Research Data Alliance formed an Interest Group tasked with developing a standardized framework for journals to use in developing such policies. (https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig)

The RDA guidance provides for several levels of policy, depending on how far a particular journal is ready to go, ranging from a minimal suggestion for data sharing to more robust requirements. The guidance lays out the elements that should be included in a data sharing policy at any level.
While few library and information science journals have developed such policies to date, it is reasonable to expect that they will, and that librarian authors will need to be prepared to comply.

In June of 2017, the Editor of the JMLA formed a working group to explore developing a policy requiring authors to share their data. The draft policy was presented to the MLA Board of Directors in May 2018 and will go into effect for manuscripts submitted on or after October 1, 2019.

http://jmla.mlanet.org/ojs/jmla/about/editorialPolicies#custom-0
JMLA Process

- Examine issues for types of data
- Query authors
- Examine other journal policies
- Track RDA project
- Iterative drafts
- Solicit input from JMLA constituencies

- Working group
  - Kevin B. Read (chair)
  - Liz Amos
  - Lisa M. Federer
  - Ayaba Logan
  - T. Scott Plutchak
  - Katherine G. Akers (JMLA EiC)

JMLA process – analyzed sampling of recent articles to examine types of data; surveyed authors to determine their willingness/ability to share data; worked to align JMLA policy with RDA guidance
JMLA Policy Summary

- Data sharing statement required
- Data to be deposited in an acceptable repository
- Materials supporting methodology are supplementary materials
- Exceptions only for proprietary or privacy concerns
- Data may be embargoed until publication
JMLA Lessons

- JMLA authors willing to share/deposit data
- Few keep data in shareable formats
- Authors need to plan for data sharing from the start
- Journal policies still evolving


JMLA lessons – most authors willing to share data, but few have data in shareable formats; journal policies are still evolving; authors need to be thinking about data sharing from the beginning of their projects.

Data Policies in Library Literature

Regina Raboin, MSLIS

Editor-in-Chief, JeSLIB

Associate Director, Lamar Soutter Library, University of Massachusetts Medical School

Data Policies in Library Literature: Regina Raboin
Before discussing best practices, it's important to understand [that] policies and procedures surrounding data are not created in a vacuum. The organizations listed above, along with others not listed here, are dedicated to providing sound ethical frameworks surrounding the use, distribution, and care of data. These organizations provide safe and informed venues for debate and discussion, fostering shared values and collaboration in order to build important and feasible data policy frameworks.

COPE recommends journals should include policies on data availability and encourage the use of reporting guidelines and registration of clinical trials and other study designs according to standard practice in their discipline https://publicationethics.org/core-practices

FORCE11 recommends data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse. In other words, data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data https://www.force11.org/datacitationprinciples | Wilkinson, M. D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 https://doi.org/10.1038/sdata.2016.18

CODATA has been taking a leadership role in developing and advancing CODATA’s Open Research Data agenda at national and international levels. The Science
International Accord on Open Data in a Big Data World, which CODATA also took the lead on drafting, recognizes both the benefits that accrue to science through open data regimes and the responsibility of scientists, and others, to make data available through appropriate means and minimal delay.  
https://zenodo.org/record/27872#.XEY1dCBO0nD

RDA: as previously mentioned by Scott, RDA’s Data Policy and Standardisation and Implementation Interest Group developed a draft standardized framework intended to assist journal editors in navigating the creation or enhancement of a data availability policy.  
https://docs.google.com/document/d/1Z4VWeQ8hMBWOpdky-Ec7GercM-4_cSDKa8gEu0eXz04/edit?usp=sharing
This slide illustrates library and information science journals varied approaches to data sharing

- The majority of the listed journals have data sharing policies, but don't require or encourage a data sharing statement.

- Interestingly, the Journal of Librarianship and Information Science offers optional open access publishing via the SAGE Choice program, but the journal Big Data & Society, also published by Sage requires a data sharing policy.

- Journal of Librarianship and Scholarly Communications offers authors their JLSC Dataverse instance to deposit and share data.

It’s important to understand that as authors you need to look for a policy at the journal level - you can’t assume that the policy will be same for all of a publisher’s journals. As a researcher and author considering how you will share data, among other things, should be considered well before writing and submitting a manuscript.
Data Sharing Challenges

- Privacy concerns, especially for clinical data
- Not being acknowledged for sharing previously
- Lack of knowledge on how to share data
- Not considered relevant (this is changing)
- It’s difficult and labor intensive
- Finding data policies


Data Availability: All data are available from Figshare at http://dx.doi.org/10.6084/m9.figshare.1288935

We want to acknowledge that data sharing compliance is not easy. Clinical and basic science researchers perceive many barriers, so it’s understandable that librarians will too. The 2015 study by Lisa Federer et al. identifies some of these obstacles.

If your research is funded, the first, and we would say the most important, priority is to be aware of your funder’s policies regarding data sharing -- some data sharing policies by funders has been compiled by SPARC at https://sparcopen.org/our-work/research-data-sharing-policy-initiative/funder-policies but if you are unable to locate a policy, you should contact your funder/program officer directly.

Always consult journal data policies before you submit, just as you would consult the author guidelines/instructions. Many library journals do not yet have a policy or guidelines, and the policies vary by journal - so you really need to go to the journal. Information is generally found under “editorial policies” or “instructions for authors”

And please note that if you have human subjects/patient data, you must also be aware of Institutional Review Board (IRB) policies and requirements and consult with this board before gathering or publishing data.
Best Practices for Data Sharing

Lisa Palmer, MSLS, AHIP

Distribution Editor, JeSLIB
Institutional Repository Librarian, Lamar Soutter Library, University of Massachusetts Medical School
Data Sharing Considerations

Thank you, Regina. Data sharing policies place expectations on authors for what to share, where to share, when to share, and how to share. For librarian researchers and all researchers, planning ahead from the start of your project helps in meeting these data sharing challenges.

This presentation doesn’t focus on “why to share” research data, as we are hopeful that the benefits of data sharing are known to our audience today. We’ll be concentrating instead on some of the practical applications of the what, when, where, and how.
What to Share

- “Minimum data set” as required by journal or funder
- Data that supports published conclusions
- Data needed to independently reproduce results
- Documentation describing the contents of the data files

**Question to Consider:**
Are there any limits to data sharing because of privacy or confidentiality?

Most journal and funder data policies stipulate that you must share publicly at least the minimum amount of data needed to independently reproduce the results and conclusions described in the manuscript, along with any supporting documentation such as a data dictionary, codebook, or readme files. We’ll go into more detail about data documentation later in the presentation.

Authors generally have some discretion as to what constitutes the “minimal data set”. For example, you usually don’t need to submit your entire data set if only a portion of the data was used in the study. In some cases it will make sense to share the raw data, while in other cases it might make sense to share manipulated or processed data.

If sharing your data is not ethical or legal -- for example, because of privacy, confidentiality, or because your Institutional Review Board has dictated that the data can’t be shared -- consult with the journal and/or funder to find out acceptable alternatives.

PLOS ONE data availability: [https://journals.plos.org/plosone/s/data-availability](https://journals.plos.org/plosone/s/data-availability)
Taylor & Francis data sharing FAQs: [https://authorservices.taylorandfrancis.com/data-sharing-faqs/](https://authorservices.taylorandfrancis.com/data-sharing-faqs/)
Journal of the Medical Library Association data sharing policy: [http://jmla.mlanet.org/ojs/jmla/about/editorialPolicies#custom-0](http://jmla.mlanet.org/ojs/jmla/about/editorialPolicies#custom-0)
When to Share

- Varies by discipline and type of data
- Typically no later than publication
- Follow guidance of journal or funder
- Embargo periods might be allowed

**Sample Data Availability Statement at time of submission:** “Data associated with this manuscript will be available in the [repository name] at [DOI].”

When you will be required to share your data will vary by discipline, type of data, and policies of the journal and funder. Typically you’ll need to provide information about where and how the data can be accessed when your manuscript is accepted, so that a DOI or other persistent identifier can be included in the published manuscript.

If *permanent* closed access is unnecessary for your data or documentation, an embargo might be a good solution. An embargo period is a formal request by an author to restrict access to documents or data for a specified period of time where the data contains, for example, sensitive information, confidential government statistics, or needs to be delayed until after all results and findings from the data have been published.

If you’re submitting a manuscript and are not yet sure exactly where you’ll be sharing your data, or your data has not yet been officially deposited in a repository, the journal may accept a placeholder statement such as the one you see on the slide: “Data associated with this manuscript will be available in the [repository name] at [DOI].”
Where to Share

- Journal website
  - Within the manuscript
  - Supplemental files
- Repository
- Recommendation: “open repository or journal”


Data Availability: All raw and processed data for this paper are shared at ScholarsArchive@OSU – Oregon State University’s repository for scholarly materials. Data may be accessed at: http://dx.doi.org/10.7267/N9W66HPQ.

Most data sharing policies require that you share your data in an appropriate public repository, unless it is already available in the submitted article.

We recommend sharing through an open repository or open access journal, for maximum access to the data and for ease of compliance with data sharing requirements. Non-repository websites -- such as departmental, lab, or personal web pages -- don’t offer the same level of visibility and aren’t as likely to sustain long-term access.

I’ll quote a sentence from the study cited on the slide, which evaluated data sharing practices of researchers at their institution: “Many of the Accessibility issues we experienced when trying to find data for this project were related to the content being shared through a closed access journal or repository.”

Journal Supplemental Files or Repository?

- Survey instruments and materials supporting the methodology are typically shared as appendices or supplemental files
- Data in journal supplemental files are not as discoverable or accessible
  - Usually do not have an individual DOI
  - Little or no metadata
  - Often in PDF format (difficult for re-use)
  - If journal is behind a paywall, access to data may be restricted to those who have subscriptions

Currently the most popular way to share data is within the journal article and supplemental files, but this is starting to change as more data repositories have become available, and journals recommend using them. For most journal data policies, the journal editor has the final decision as to where data is made available, following the norms the journal has established.

Materials supporting the methodology -- such as survey instruments, questionnaires, rubrics, assessment instruments, scales, and figures -- are typically shared as appendices or supplemental files on the journal website. This is reflected in Journal of the Medical Library Association’s policy, as Scott explained earlier.

It’s important to point out that journal supplemental files often do not have their own DOIs or much metadata, making them more difficult to find and cite. This is especially true if you are publishing in a subscription journal, where access to data may be restricted and closed to those without subscriptions.
Repository Options

- Disciplinary or domain
- Institutional
- General (e.g. Figshare, Open Science Framework, Zenodo, Dryad, Harvard Dataverse, OpenICPSR)

As a general rule, a domain repository is likely to offer the best home for data than can be publicly shared, because data is available to your peers in a known, authoritative place. For library science, there really isn’t a disciplinary data repository at this point, although your data may be appropriate for one in another discipline. But institutional or general repositories are great solutions for library research data.

The table on the slide is from a paper by Lisa Federer and her colleagues that analyzed the Data Availability Statements from more than 47K papers published in PLOS ONE between March 2014 and May 2016. This table lists the 20 most frequently mentioned repositories or sources. Many of these sources are ones that librarians might utilize -- and are utilizing -- such as Figshare, Dataverse, the Open Science Framework, Zenodo, and of course, institutional repositories.


Data Availability: The final dataset and accompanying code are available on the Open Science Framework, https://doi.org/10.17605/OSF.IO/UN8JX
Choosing a Repository -- Criteria to Consider

- Journal recommendations
- Local repositories available to you
- Reputation
- Sustainability
- Visibility (e.g. DOI or other unique identifier)
- Usability
- Features (e.g. embargoes, author dashboard, Creative Commons licensing)
- Formats (e.g. can it take your data)
- Rights

How do you choose a data repository? Here are some criteria to consider. Make sure to consult any recommendations that the journal makes for data deposit and look into any local repositories that might be at your disposal, such as at your institution or available through an institutional membership. You want the repository to be reputable and to be backed by an institution, community, or funder. Make sure you can obtain a DOI or other unique identifier, which will make it easier for your data to be discovered and cited. Features to look for include the ability to set an embargo, track usage data, and apply a Creative Commons license. Make sure the repository can accommodate the data you've generated and gives you sufficient rights.


Boston University Data Services: What is a Data Repository? [https://www.bu.edu/data/share/selecting-a-data-repository](https://www.bu.edu/data/share/selecting-a-data-repository)
There are several registries and other sources you can consult to help you address those considerations and select an appropriate data repository. R3data.org and FAIRsharing.org help you discover a reputable repository, as does the new Repository Finder tool hosted by DataCite. The Digital Curation Centre has a checklist that addresses important criteria, and MIT provides a template to compare repositories. Harvard Medical School’s matrix compares several of the general data repositories.

- Registry of Research Data Repositories: https://www.re3data.org
- FAIRsharing: https://fairsharing.org/
- Repository Finder: https://repositoryfinder.datacite.org/
- MIT Libraries Repository Comparison Template: https://libraries.mit.edu/data-management/share/find-repository
- HMS Biomedical Data Repositories Matrix: https://datamanagement.hms.harvard.edu/repositories
A journal may also make a data repository available to its authors. For example, the Journal of Librarianship and Scholarly Communication has its own instance of Dataverse. The Journal of eScience Librarianship’s data policy is still evolving, but because the journal platform is also a repository, we plan to offer authors the option to deposit their data with the journal. Benefits include the ability to provide a DOI for the dataset (separate from the article), Creative Commons licenses, usage metrics, and embargoes.

Now I’ll turn the presentation over to Julie Goldman to highlight some best practices.

Open data in eScholarship@UMMS: https://libraryguides.umassmed.edu/openaccess/open_data#s-lg-box-10569427

Journal of Librarianship and Scholarly Communication data sharing policy: https://jlsc-pub.org/about/editorialpolicies/#data-sharing

Author guidelines: https://jlsc-pub.org/about/submissions/#data
Best Practices for Data Sharing

Julie Goldman, MLIS

Managing Editor, JeSLIB

IResearch Data Services Librarian, Countway Library, Harvard Medical School
Best Practices

- Plan for data sharing at the start of your research project
- Follow well-structured, standardized data entry and validation
- Use open and non-proprietary file formats
- Provide documentation for your project and data
- Share completely de-identified data
- Include data availability statement with a permanent unique identifier
- Cite any and all datasets you use

So now that we have looked at the what, when and where to share, let's discuss how to approach data sharing. Here are some best practices for preparing your data for sharing. These are just highlights, and we are taking a “practice what we preach” mentality.

- Again, we want to stress that planning ahead from the start of your research project helps in meeting these data sharing challenges. You should include what data you plan to share in your original research process - it is always a good idea to create a data management plan!
- You should follow well-structured, standardized data entry and validation. Much of the research process is spent on the ‘data wrangling’ stage, but some of it can be prevented with good strategies for data collection up front.
- Use open and non-proprietary file formats. In order to maximize accessibility and ensure long term preservation, we encourage the use of non-proprietary formats. Spreadsheets are preferable to PDF for tabular data, so using comma-delimited text files (.csv) or plain text files (.txt) rather than Excel (.xls, .xlsx) and Word (.doc, .docx) documents are the best choices.
- Provide documentation for your project and data. It is a good practice to begin to document your data at the very beginning of your research project and continue to add information as the project progresses.
- Share completely de-identified data. This refers to the process of removing or
• obscurring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them.

• Include a data availability statement with a permanent unique identifier. This is also sometimes referred to as a 'data access statement' and is crucial in signposting where the data associated with a paper is available, and under what conditions the data can be accessed, including links to the data set.

• Finally, cite any and all datasets you use, including your own. While citation standards for data sets are still evolving, it is only best scholarly practice to cite any and all sources you use, and this includes data. These last two points help enhance discoverability of your and other’s work.

Now we'll go into more detail on a few of these and share some helpful resources.
Data Entry & Validation

- Make sure your data is well-structured, standardized, and labeled

- Use standardized date formats such as YYYYMMDD


- Utilize data validation for data collection and entry

- Use a data collection tool such as REDCap to ensure your data is structured appropriately

- Structured data has been organized so that its elements can be made accessible for more effective analysis. I highly suggest the Library Carpentry lesson ‘Tidy Data for Librarians’, which provides tips and tricks for working in Excel and OpenRefine, and ensuring you are working with tidy, well-structured and labeled data.
  
  ○ Library Carpentry, Tidy Data for Librarians: [https://librarycarpentry.org/lc-spreadsheets](https://librarycarpentry.org/lc-spreadsheets)

- Apply standardized date formats such as ISO 8601. This standard provides needed consistency in date formatting, allows for inclusion of several types of date-time information, and can sort dates chronologically. If you have not read Kristin Briney’s JeSLIB commentary on applying this standard to your everyday work, we highly suggest it. This is a simple action you can start applying to your daily practice today!
  

- When you have a well-structured data table, you can use several simple techniques within your spreadsheet to ensure the data you enter is free of errors. These approaches include techniques that are implemented prior to entering data (quality assurance) and techniques that are used after entering data to check for errors (quality control). Again, there are many tips in the
- Library Carpentry lessons you can explore here.
  - Library Carpentry, Tidy Data for Librarians: [https://librarycarpentry.org/lc-spreadsheets](https://librarycarpentry.org/lc-spreadsheets)
- We recommend using REDCap for data collection over other survey collection tools. REDCap is a free, web-based, and user-friendly electronic data capture tools, useful for collecting and tracking information and data from research studies. It really is a data management and survey tool in one, because, just to name a few features, it is fully customizable, has multi site access, data import functions, you can export survey results to common data analysis packages, and is a secure, HIPAA-compliant option.
  - REDCap: [https://projectredcap.org](https://projectredcap.org)
Documentation

- Metadata about the project should be conveyed in a simple README file
- Provide a data dictionary or codebook to explain column names and other information in spreadsheets
- REDCap generates a data dictionary based on your data collection forms
- Consult experts and available resources

- Data documentation is the key for future understanding of research data. Along with your data, you should provide documentation (“metadata”) that helps you and everyone on your project team understand your data, and also helps other researchers find, use, and properly cite your data. Metadata “completes” a dataset.
  - At a minimum, create a readme text file that includes basic documentation such as title, creator, identifier, rights and access information, dates, location, methodology, etc.
  - Ideally, you want to use standard terminology to enable others to find and use your data. Identify keywords and ontologies for your area of research, if they are available.
- In REDCap, a Data Dictionary is created in CSV format representing the structure of your database. It contains the metadata used to construct your data collection instruments. In addition a human-readable version is accessible, known as the "Codebook", which serves as a quick reference that lets you view the attributes of any given field in the project without having to download and interpret the Data Dictionary. Therefore eliminating steps for you to create these items yourself!
- We acknowledge that we have already covered a lot, and even following the minimum standards for sharing can be difficult and time consuming! So obviously we encourage you to consult experts and resources for assistance
and advice, and we have included many resources throughout the slides and a full list at the end.

  
  - “We recognize that data sharing and meeting even the minimum standards for sharing [...] can be difficult and time consuming, and we encourage authors to engage with research data service providers either at the local level (often part of their academic library) or at the national level (e.g. DataONE — [www.dataone.org](http://www.dataone.org) DataQ — [www.researchdataq.org](http://www.researchdataq.org) etc.) for assistance and advice.”

Additional documentation resources:

- Kristin Briney provides an excellent overview of data documentation in her book, which is listed at the end on the Resources slide.
- Guide to writing “readme” style metadata (Cornell University): [https://data.research.cornell.edu/content/readme](https://data.research.cornell.edu/content/readme) (also available as a pdf [http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf](http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf) with example readme files) and readme file template ([https://cornell.app.box.com/v/ReadmeTemplate](https://cornell.app.box.com/v/ReadmeTemplate))
- DataONE Best Practices Primer: [https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf](https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf)
Data De-Identification

- Hide **direct** confidential identifiers of people and organizations, e.g. names, geographic information, dates, telephone numbers, email addresses, URLs

- Hide **indirect** identifiers that when linked with other available information could identify someone

De-identification of data refers to the process of removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. It is considered successful when there is no reasonable basis to believe that the remaining information in the records can be used to identify an individual.

- You should hide **direct** confidential identifiers of people and organizations, e.g. names, geographic information, dates, telephone numbers, email addresses, URLs — any unique identifying number, characteristic, or code.

- Also, hide indirect **identifiers** that when linked with other available information could identify someone or their institution.

- Approaches include replacing personal information with surrogates that can later be used to look up the real values, dropping the columns or recoding the variables.

See these two tools as options for de-identifying data:

- **NLM-Scrubber** for unstructured text - which is out of the National Library of Medicine.

- **OpenRefine** for structured data - which is open source and I highly suggest librarians get familiar with!
  - OpenRefine: [http://openrefine.org](http://openrefine.org)
I'll also mention REDCap again as it is more secure option than Microsoft Excel. It is also HIPAA-compliant, as fields in REDCap can be marked as identifiable; and the user has the option of de-identifying their data during export. REDCap also offers daily backups, basic support, and an audit trail feature for even more security.

Additional resources:

- “Preparing data for deposit” from UK Data Archive: [https://ukdataservice.ac.uk/deposit-data/preparing-data.aspx](https://ukdataservice.ac.uk/deposit-data/preparing-data.aspx) and “Depositing shareable survey data” [https://ukdataservice.ac.uk/media/440320/depositssurvey.pdf](https://ukdataservice.ac.uk/media/440320/depositssurvey.pdf)
- “Preparing data for sharing” from ICPSR: [https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html](https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter5.html)
- How to anonymize quantitative data (UK Data Archive): [https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/quantitative](https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/quantitative)
- How to anonymize qualitative data (UK Data Archive): [https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative](https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative)
Enhancing discoverability to your data involves including a data availability statement with a permanent unique identifier, allowing others to successfully cite your work.

The first example on the slide is from Steven Van Tuyl and Amanda Whitmire’s 2016 evaluation of data sharing practices by OSU researchers funded by the NSF. The figure shows low scores in discoverability and accessibility to data. These low scores stem from issues surrounding standards in data citation styles and methods -- issues in the fact that they varied widely -- making it difficult to find and access data that was meant to be shared.

  - “One of the biggest Discoverability problems we saw with data sharing through journal articles was that citation styles and methods varied widely, making it difficult to find data that was meant to be shared.”
  - Also, the authors followed best practices for data citation by citing their dataset in this paper.

The second example is from Lisa Federer et al.’s 2018 research of data availability statements from over 47,000 papers published in PLOS ONE between March 2014 and May 2016. The figure shows ‘articles missing a Data Availability Statement over
time’ (the red line indicates total published articles, while the bars indicate articles with no Data Availability Statement). While we can see that once PLOS’ data sharing policy was implemented in March 2014 the amount of articles missing a data availability statement decreased, within their study the authors still found some inconsistency with the statements that were included. For example, where a direct URL, DOI, or accession number was not included, basically making it very hard for others to actually directly navigate to or find the data tied to the article. I’ll also point out that Lisa’s article also has an image which shows sample statements, which might be helpful for people.


  - “18.2% of PLOS papers named a specific repository or source where data were available. Most Data Accessibility Statements direct the reader to the paper itself or supplementary information. Even among the data repository articles, some Data Accessibility Statements indicated a repository but failed to include a URL, DOI, or accession number — basically sending readers on a wild goose chase to locate their data within the repository.”

We'll finish up by exploring some examples of data deposit from JeSLIB & JMLA.

This first example is from JeSLIB where a data file was submitted along with the article as supplemental material. There is no data availability statement, since the data file has been submitted as a supplementary file hosted as part of the article record on our eScholarship@UMMS platform. We are sharing this example, because as we think about revamping our own data policy at JeSLIB, we probably want to state in our policy that specific file formats should be submitted, such as csv over those proprietary file formats we discussed earlier.


Supplemental Content: Data File | An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2016.1089 under “Additional Files”.
This is a second example in JeSLIB, where data supporting a submitted article was made available in an institutional repository. So we have included a nice Data Availability statement that points to the DOI of the dataset deposited in the Illinois Data Bank. The statement is found both within the article text, as well as within the metadata on the article landing page.

https://doi.org/10.7191/jeslib.2017.1112

Data Availability: Data collected for this study is available in the Illinois Data Bank at https://doi.org/10.13012/B2IDB-1235375_V1
The most recent example of data deposit we can share from JeSLIB, shows where data supporting a submitted article was made available in the general repository, Zenodo. Again, we have included a nice Data Availability statement that points to the DOI of the dataset in Zenodo. The author included a data citation for this dataset in the manuscript.


Data Availability: Data associated with this article are available from Zenodo at https://doi.org/10.5281/zenodo.2432419
This final example is from JMLA where data, including code and analysis, have been made available via the Open Science Framework. We note here that no ‘Data Availability’ statement is provided, instead a sentence is included in the ‘Data analysis’ section of the Methods, pointing to the globally unique identifier (GUID) for that OSF project. This is because the JMLA Data Policy has not gone into effect yet; if this article were to be published come October, 2019, as Scott described earlier, a structured ‘Data Availability’ statement would be included.


METHODS
Data analysis: The full code for this analysis and the de-identified dataset are available on Open Science Framework: https://osf.io/zafr6.

- Dataset includes: README file; Data, in csv file (illustrates how you can make your survey data available and also redact survey comments to protect privacy of respondents); R code used for analysis; article also has 2 PDF appendices as supplementary materials (the survey instrument and the taxonomy)
Nine Simple Ways

1. Share your data.
2. Provide metadata.
3. Provide an unprocessed form of the data.
4. Use standard data formats.
5. Use good null values.
6. Make it easy to combine your data with other data sets.
7. Perform basic quality control.
8. Use an established repository.
9. Use an established and open license.


To close, here are “Nine Simple Ways” to make your data understandable, easy to analyze, and readily available to the wider community, applicable to both scientists and librarians.

1. Share your data.
2. Provide metadata. -- we discussed documentation
3. Provide an unprocessed form of the data. -- sharing the raw data, unanalyzed or code
4. Use standard data formats. -- we mentioned a few standards such as ISO 8601 for dates
5. Use good null values.
6. Make it easy to combine your data with other data sets. -- this is where tidy, well-structured data will be easy to work with
7. Perform basic quality control.
8. Use an established repository. -- Lisa described qualities to look for when evaluating a repository to choose
9. Use an established and open license.

Questions?

Regina Raboin, MSLIS
Editor-in-Chief, JeSLIB
UMass Medical School

Lisa Palmer, MSLS, AHIP
Distribution Editor, JeSLIB
UMass Medical School

Julie Goldman, MLIS
Managing Editor, JeSLIB
Harvard Medical School

T. Scott Plutchak, MA, AHIP
Editorial Board, JeSLIB & JMLA
University of Alabama–Birmingham (Retired)

Contact Us: https://escholarship.umassmed.edu/jeslib/editorialboard.html

Contact Us: https://escholarship.umassmed.edu/jeslib/editorialboard.html
Additional Resources


23 Things: Libraries for Research Data: http://dx.doi.org/10.15497/RDA00005


Library Carpentry Lessons: https://librarycarpentry.org/lessons

Additional Resources - See also the slide notes for more specific resources

- 23 Things: Libraries for Research Data: http://dx.doi.org/10.15497/RDA00005
- Library Carpentry Lessons: https://librarycarpentry.org/lessons
Funding Statement

JeSLIB has been partially funded by the National Library of Medicine, National Institutes of Health, Department of Health and Human Services, initiated under Contract No. HHS-N-276-2011-00010-C with the University of Massachusetts Medical School, and continues with partial funding from Cooperative Agreement Number UG4LM012347 with the University of Massachusetts Medical School.