This is an original manuscript of an article published by Taylor & Francis in *Journal of Hospital Librarianship* on January 2024 available at: https://doi.org/10.1080/15323269.2024.2326787

Exploring Freely Available Data Tools to Support Open Data and Open Science

Christine Nieman Hislop[a], Katie Pierce Farrier[b], Elizabeth Roth[c]

[a]*Health Sciences and Human Services Library, University of Maryland, Baltimore, Baltimore, USA;* [b]*Gibson D. Lewis Library, University of North Texas Health Science Center, Fort Worth, USA;* [c]*James W. Colbert Education Center and Library, Medical University of South Carolina Libraries, Charleston, USA.*

Christine Nieman Hislop is the Data Education Librarian for the Network of the National Library of Medicine (NNLM) Region 1 at the University of Maryland, Baltimore. https://orcid.org/0000-0002-6844-4228, cnieman@hshsl.umaryland.edu

Katie Pierce Farrier is the Data Science Strategist for NNLM Region 3. https://orcid.org/0009-0006-8559-0223

Elizabeth Roth is the Research and Data Science Strategist for NNLM Region 2. https://orcid.org/0009-0005-6326-5887

Citation:

Hislop, C. N., Farrier, K. P., & Roth, E. (2024). Exploring Freely Available Data Tools to Support Open Data and Open Science. *Journal of Hospital Librarianship, 24*(2), 104–111. https://doi.org/10.1080/15323269.2024.2326787

Exploring Freely Available Data Tools to Support Open Data and Open Science

Librarians support researchers by promoting open science and open data practices. This article explores five freely available tools that support and facilitate open science practices. Open Science Framework provides a platform for project management, data sharing, and data storage. OpenRefine cleans and formats data. DMPTool has templates for data management and sharing plans that comply with funder mandates. The NIH Common Data Elements is a repository for standardized data elements, and finally, the NLM Scrubber is a tool for de-identifying clinical text data. Information professionals can add these tools to their repertoire and share them with researchers at their institution.

Keywords: open data, open science, data management

Acknowledgements: This work was supported by the National Library of Medicine (NLM), National Institutes of Health (NIH), under Cooperative Agreement UG4L013724, UG4L013736, and UG4LM012345. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Open Data and Open Science

Open data, open science, open education, open access, open everything! The movement for more collaboration, information, and freedom from paywalls continues to grow. However, what do all these “opens” mean? Open data is data that is freely available for reuse or redistribution, preferably easy to access (such as downloading over the internet), and with few restrictions, such as attribution and Share Alike (1). Open data also means FAIR data, which means data is findable, accessible, interoperable, and reusable data (2). Not all data can or should be shared so openly, but steps toward such openness significantly contribute to advancing science and knowledge as a whole.

While open data is specific to data, open science refers to the broader process of generating knowledge. Foster Open Science defines open science as

> "the practice of science in such a way that others can collaborate and contribute, where research data, lab notes, and other research processes are freely available, under terms that enable reuse, redistribution, and reproduction of the research and its underlying data and methods." (3)

Open science touches every stage of the research lifecycle, so open data and open access fall under the larger umbrella of open science.

Impact and Importance of Open

Open data benefits scientific advancement and society as a whole, but there are also funding and publisher mandates that call for sharing data when possible. For example, NIH strongly encourages researchers to share data whenever possible. In early 2023, the NIH issued a data management and sharing (DMS) policy requiring NIH-funded researchers to develop a 2-page plan outlining the management, sharing, and preservation of research data (4).

Furthermore, open science and open data promote reproducibility and transparency. Transparency is when scientists publish their methodologies to provide evidence of merit, whereas reproducibility refers to the ability of others to reproduce findings and generate new data that supports the same conclusions (5). Both are crucial to promoting good science. An investigation into the reproducibility of cancer biology research found that only 4 out of 193 experiments had enough data publicly accessible to compute, and none of the experiments had enough details available to redesign the experiment protocols (5). Using data tools that support open science can help mitigate barriers to transparent and reproducible science and help researchers comply with funder mandates.

The following five freely available tools, Open Science Framework, OpenRefine, DMPTool, NIH Common Data Elements, and NLM Scrubber, all support data management and sharing by providing ways to manage, wrangle, clean, organize, and de-identify data. Not every researcher will need to use every single one of these tools, but they can help researchers better manage their data and make it easier to comply with NIH and publisher policies.

Open Science Framework (OSF)

The Open Science Framework (OSF) is a free, open platform for organizing, managing, and sharing research projects that support and promote collaboration (6). Created and hosted by the Center of Open Science (COS), OSF supports researchers throughout a project's lifecycle. An account is free, and users can register projects, document research procedures, find and work with collaborators, and share preprints or research data once the project is complete. Projects in OSF are private unless shared with select collaborators.

Uses of OSF

OSF has a breadth of features that support open science and open data. OSF connects with other tools like cloud servers (GitHub, Google Scholar), citation managers (Zotero, Mendeley), and scholarly identifiers, such as a DOI, to facilitate more open science at every stage of research. Users can search or browse through OSF's public projects to identify collaborators and share or review preprints of the latest research.

To support the implementation of the NIH Data Management and Sharing Policy, the NIH has partnered with six generalist repositories through the General Repository Ecosystem Initiative; these repositories include OSF, Dryad, Dataverse, Figshare, Mendeley Data, and Vivli (7). These repositories are working to develop

consistent metadata to maximize discoverability and interoperability between repositories. Files, including datasets, text files, CSVs, and other data, can be stored in OSF up to 2 GB, and data files or whole projects can be shared with collaborators on the platform or made publicly available. Data deposited in OSF can be shared with collaborators alongside documentation such as a README or data dictionary, increasing the accessibility and reusability of the research data (6, 8).

OSF is more geared toward hard sciences rather than social sciences and clinical research, but its collaboration options may still be of interest. Librarians can use OSF to house their own research data or collaborate with researchers on various projects. Rather than trying to keep track of multiple disparate tools, OSF can act as a unifying tool because it connects researchers to repositories, citation managers, and cloud servers.

OpenRefine

OpenRefine is a free, open source tool that can clean and transform messy data (9). For example, users can make mass edits to columns and values to fix spelling and formatting. Researchers and librarians can use OpenRefine to prepare and modify datasets for analysis and visualization programs (10). OpenRefine works particularly well with larger datasets or numerous data files that need to be cleaned in the same way.

Uses of OpenRefine

Developed as a Google project, OpenRefine is now community-driven (9, 10). It uses a web browser as an interface, but the program (and data) is not connected to the internet. Users can safely work with sensitive data while maintaining privacy. OpenRefine has several built-in powerful algorithms that look for and cluster similar phrases or word patterns to make mass edits and formatting changes (10). The program tracks and saves

any changes to the dataset. Users can save processing commands and reuse them with future datasets. For larger datasets, reproducing these processing steps saves users' time (by not repeating individual steps multiple times on multiple files) and ensures consistency and replicability. This documentation is especially useful for complying with the NIH DMS policy. OpenRefine preserves the dataset's metadata, and future researchers can replicate data cleaning and formatting steps.

While OpenRefine has many features, the interface can be intimidating and has a learning curve. OpenRefine is best suited for large or multiple datasets that need to be cleaned and formatted using the same steps. However, ample documentation is available on OpenRefine's website. Users can find guidance on using facets, transposing columns, and transforming data using regular expressions (11). In addition to its code being open source, all of OpenRefine's help documentation is openly licensed for reuse.

DMPTool

The DMPTool is an online application that helps researchers create data management plans that comply with funder mandates, most notably the NIH DMS policy (13). The DMPTool provides guidance on best practices for data management and template plans. DMPTool fits in at the very start of the research cycle because it helps lay the initial groundwork and helps researchers consider making their data more open and FAIR before the bulk of the research even starts. Anyone can create an account and create plans for free. Some organizations have institutional accounts that allow them to direct users to internal resources and data services.

The site provides step-by-step guidance for filling out each template section and identifies data management and sharing considerations researchers should address when facilitating a large, federally funded project. DMPTool created templates tailored to meet different funder requirements, such as the NSF, Department of Defense, or USDA

(14). To prepare and comply with the NIH DSM policy requirements, a special working group created sample language templates to help researchers write more comprehensive data management and sharing plans (14).

Uses of DMPTool

The DMPTool is a wonderful addition to outreach and instruction. The site provides abundant outreach materials, including talking points, logos, customizable slide decks, and general data management guidance. Librarians can easily modify the provided materials to their specific institution and use the content for instruction and outreach to researchers. The DMPTool is user-friendly and allows users to create mock data management plans. Researchers can walk through the steps, explore the website's resources, and practice writing data management plans. The tool also promotes collaboration and feedback. Researchers can share drafts and completed plans within their institution or add collaborators or editors to review a plan. Users can also browse publicly available DMPs and see examples of completed plans, such as this one on brain growth and congenital heart disease (15).

NIH Common Data Elements

A common data element (CDE) is a standardized, precisely defined question (i.e., variable) paired with specific rules for allowable responses (16). CDEs may be used in data collection instruments across multiple datasets, research sites, studies, etc. Data that is collected, recorded, and organized in consistent ways makes datasets easier to share, analyse, and reuse. CDEs support open data and open science because they provide transparency in how data is collected and standardized so that data can be more easily reproduced or potentially reused for other research endeavours.

The NIH CDE Repository is a repository of common data elements recommended or required by NIH Institutes and Centers (16). The CDEs in the NIH CDE Repository provide standards for how medical questionnaires or surveys are structured so that questions and responses are captured consistently across different studies, organizations, and disciplines. NIH CDEs offer commonly used questions and answer possibilities. Some CDEs have been reviewed and endorsed by the NIH CDE Governance Committee. The committee reviews CDEs for clearly defined variables, documented reliability, human and machine readability, whether it is recognized or recommended by an NIH Institute or Center, and intellectual property licensing (16). Any CDEs found in the NIH CDE Repository can contribute to more standardized, reproducible data practices.

Use of NIH CDE

Librarians can help promote NIH CDE through presentations and outreach efforts. The new NIH DMS policy asks researchers what data or metadata standards they plan to use in their research. Researchers can use the NIH CDE Repository to find commonly used forms or elements to aid data collection methods. These standardized measures can reduce the time needed to normalize and clean data. For example, CDEs allow for comparisons across cohorts and help facilitate more extensive studies (17).

NIH CDE also has an online tutorial that reviews the FAIR (Findable, Accessible, Interoperable, Reusable) data principles and shows learners how CDEs can support the creation of FAIR data (18). The site offers robust training and help resources, including webinars and blog posts introducing common data elements, how they are used, and how to effectively use the NIH CDE Repository (18). Researchers and librarians are encouraged to take these free trainings and learn more about this

resource. However, the NIH CDE Repository is not comprehensive or universal. Certain forms and elements may differ across fields and institutions. Researchers should take care to use the CDEs that best suit their needs.

NLM Scrubber

Developed at the Lister Hill National Center for Biomedical Communications, the NLM-Scrubber is a freely available tool that uses natural language processing (NLP) to find personally identifying information (PII) and de-identify it (18). Natural language processing is a branch of artificial intelligence that focuses on understanding human languages. NLP uses human languages, statistics, and machine learning to train computers to understand the intent, meaning, and sentiment of voice or text data (19). For the NLM Scrubber, the program is trained to recognize PII and replace it with generic, non-identifying terms per the Health Insurance Portability and Accountability Act of 1996 (HIPAA). For example, "John Smith, diagnosed at age 92," is replaced with "[PERSONALNAME], diagnosed at age [AGE90+]." However, even the most advanced computer programs are not infallible, and a human should verify that any sensitive information is truly de-identified to HIPAA standards before sharing.

Once clinical notes or electronic medical records are de-identified, data can be shared or analysed more safely. NLM Scrubber supports open science and open data because once data has been de-identified to HIPAA standards, it can be shared and reused (assuming participants consented to the sharing and reusing their data). Clinical notes provide a trove of health information that may benefit multiple research projects and fields.

Uses of NLM Scrubber

Librarians can promote NLM Scrubber to researchers and faculty involved in clinical

research projects. Researchers can customize the list of terms to be targeted and replaced by the NLP program. Researchers need to reformat files into flat text files before de-identification, but it runs faster than comparable products such as ClinDeID and Amazon Comprehend (20). NLM Scrubber is available for Windows and Linux systems and will never be monetized since a federally funded research project developed it. Overall, NLM Scrubber de-identifies clinical text data so that it can be more safely shared and saved. Librarians can incorporate NLM Scrubber as a resource when discussing clinical research and promoting open data practices. An important note: NLM Scrubber is not fool proof and may not capture all PII (21). Best practices require a human element to verify de-identification and anonymization before sharing potentially sensitive data.

Conclusions

Open science and open data continue to play a vital role in building trust, transparency and furthering scientific knowledge. Librarians can support researchers by promoting open science and open data practices. Librarians can bring more awareness to the importance of open data and open science. We encourage librarians to incorporate these freely available tools into their outreach, instruction, and research consultations. NLM Scrubber and NIH Common Data Elements, in particular, are less well known but could greatly impact the ability to share, reuse, and reproduce research data.

Product guides, detailing popular uses, key points, and more information resources, are available for each of these tools.

References

(1) Dietrich, D, Gray J, McNamara T, Poikola A, Pollock R, Tait J, and Zijlstra T. What Is Open Data? [Internet]. USA: Open Data Handbook; 2023 [cited 2023

Aug 7]. Available from https://opendatahandbook.org/guide/en/what-is-open-data/.

(2) FAIR Principles [Internet]. USA: Go FAIR; 2023 https://www.go-fair.org/fair-principles/

(3) What is Open Science? [Internet]. USA: FOSTER; 2023 Available from: https://www.fosteropenscience.eu/learning/what-is-open-science/#/

(4) Data Management and Sharing Policy [Internet]. Bethesda, MD, USA: National Institute of Health; 2023. Available from: https://sharing.nih.gov/data-management-and-sharing-policy

(5) Errington T, Iorns E, Gunn W, Tan F, Lomax J, and Nosek B. An Open Investigation of the Reproducibility of Cancer Biology Research [Internet]. 2014 Dec 10. Available from: https://doi.org/10.7554/elife.04333.

(6) Open Science Framework [Internet]. OSF Home. USA: The Center for Open Science; 2023. Available from: https://osf.io/

(7) Generalist Repository Ecosystem Initiative [Internet]. Bethesda, MD, USA: NIH Office of Data Science Strategy; 2023 May 26. Available from: https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative

(8) Foster ED, Deardorff A. Open Science Framework [Internet]. 2017 Apr 4 [cited 2023 Sep 7]. Available from: https://doi.org/10.5195/jmla.2017.88

(9) OpenRefine. [Internet]. USA: OpenRefine; 2023. Available from: https://openrefine.org/

(10) Hill KM. In Search of Useful Collection Metadata: Using OpenRefine to Create Accurate, Complete, and Clean Title-level Collection Information [Internet].

2016 Sept 14 [cited 2023 Aug 31]. Available from: http://dx.doi.org/10.1080/00987913.2016.1214529

(11) OpenRefine User Manual [Internet]. USA: OpenRefine; 2022. Available from: https://openrefine.org/docs

(12) DMPTool: About [Internet]. CA, USA: California Digital Library; 2023. Available from: https://dmptool.org/about_us

(13) DMP Templates: Funder Requirements [Internet]. CA, USA: California Digital Library; 2023 Available from: https://dmptool.org/public_templates

(14) Praetzellis M. Updates on DMPTool Support for the NIH DMSP Requirements. [Internet]. CA, USA: DMPTool Blog; 2023 Jan 24. Available from: https://blog.dmptool.org/2023/01/09/updates-on-dmptool-support-for-the-nih-dmsp-requirements/

(15) Ortinau, C. Effects of Placental Dysfunction on Brain Growth in Congenital Heart Disease [Internet]. CA, USA: DMPHub; 2013 Mar 16 [Updated 2023 Apr 21]. Available from: https://dmphub.cdlib.org/dmps/doi:10.48321/D1BS7N

(16) Guide to the NIH CDE Repository. [Internet.] Bethesda MD, USA: National Library of Medicine; 2023. Available from: https://cde.nlm.nih.gov/guides#introduction

(17) Dennis E, Finian K, Tate D, and Wilde E. The Role of Neuroimaging in Evolving TBI Research and Clinical Practice [Internet]. 2023 Feb 26 [cited August 2 2023]. Available from: https://doi.org/10.1101/2023.02.24.23286258.

(18) Resources: CDE Trainings from NLM. [Internet]. Bethesda, MD, USA: National Library of Medicine; 2023. Available from: https://cde.nlm.nih.gov/resources

(19) Lister Hill National Center for Biomedical Communications. Clinical text de-identification using NLM-Scrubber. Bethesda MD, USA: National Library of Medicine; 2023. Available from: https://lhncbc.nlm.nih.gov/scrubber/

(20) What is natural language processing? [Internet]. USA: IBM; 2023. Available from: https://www.ibm.com/topics/natural-language-processing

(21) Heider P, Obeid J, and Meystre S. A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools. [Internet]. 2020 May 30 [Cited 2023 Sept 7]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/pmc7233098/

(22) Steinkamp J, Pomeranz T, Adleberg J, Kahn C, and Cook T. Evaluation of Automated Public De-Identification Tools on a Corpus of Radiology Reports. [Internet]. 2020 Oct 14 [Cited 2023 Sept 7] Available from: https://doi.org/10.1148/ryai.2020190137