Research Data Management Plan for the Meaningful Data Counts Project

Anton Ninkov[1,2], Kathleen Gregory[1,2,3], Chantal Ripp[2,4], Erica Morissette[1,2], Lina Harper[1,2], Isabella Peters[5,6], Felicity Tayler[4], & Stefanie Haustein[1,2,7,*]

[1] School of Information Studies, University of Ottawa, Ottawa (Canada)
[2] Scholarly Communications Lab, Ottawa/Vancouver (Canada)
[3] University of Vienna (Austria)
[4] Library, University of Ottawa, Ottawa (Canada)
[5] ZBW Leibniz Center for Economics, Kiel (Germany)
[6] Kiel University, Kiel (Germany)
[7] Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montreal, (Canada)
[*] stefanie.haustein@uottawa.ca

Version	Rationale	Authors
1.1	Originally included in v.1 of the RDMP to build awareness of costs associated with RDM activities, itemized resources were removed and are managed via institutional grant management system.	S. Haustein; C. Ripp
1.2	Minor changes reflected to Data Collection section (A4): bibliometrics data collected from DataCite and Dryad; method of sampling researchers for the survey based on institutional emails from Web of Science; and the specification of the tool NVivo for coding interview transcripts.	S. Haustein; A. Ninkov
1.3	Minor changes reflected to Data Collection section (A5) regarding: file formats of recordings (.MP3, .MOV) and .NVPX for NVivo projects.	A. Ninkov; K. Gregory
1.4	Changes to Data Collection section (A6) detailing documentation and file naming practices (e.g., using YYMMDD in the file name rather than v1, v2). Using date and descriptor in the file labeling allows project team members to reflect on when a new file was created and for what purpose.	K. Gregory
1.5	Clarification on practices in the Storage and Backup section (A13) related to removal of participants who did not want to be included in the survey sample.	A. Ninkov
2.0	Introductory statement specifies which stage of the project the RDMP reflects.	C. Ripp S. Haustein

About the RDMP

This research data management plan (RDMP) describes how data is handled in the Meaningful Data Counts research project. The project is led by PI Stefanie Haustein and Co-PI Isabella Peters and funded by the Alfred P. Sloan Foundation. It uses mixed social sciences research methods, producing quantitative as well as qualitative datasets. As both PIs practice open science and scholarship, the RDMP puts an emphasis on data sharing and reuse. All data, software code and milestone working documents will be made available freely in online repositories throughout the research project and all publications will be published open access. The RDMP is structured and based on a template created for the Portage Network (new Digital Research Alliance of Canada - the Alliance). The RDMP is a living document that has been revised as the

project has evolved. The version control table reflects the important changes to the sections of the RDMP. The culmination of changes leading to version 2 of the RDMP occurred during the middle stage of the project, following data collection of the bibliometrics data and survey responses, allowing the project team to validate and reflect current data management practices.

RESPONSIBILITIES AND RESOURCES

Q1. Who will be responsible for data management? Will the Principal Investigator (PI) hold all responsibility during and beyond the project, or will this be divided among a team or partner organizations?

A1. The PI is responsible for all research data. She set up necessary infrastructure and makes sure that all team members are aware of individual responsibilities from the beginning of their involvement in the project. Should the PI leave the project, the co-PI will take responsibility. Although the main responsibility rests with the PI and Co-PI, every team member, particularly lead authors, will be responsible for proper documentation, versioning, storage and publication of research data as part of their usual workflow.

Guidance: Succession planning. The PI is usually in charge of maintaining data accessibility standards for the team. Consider who will field questions about accessing information or granting access to the data in the event the PI leaves the project. Usually, the Co-PI takes over the responsibilities.

Q2. In the event that the PI leaves the project, who will replace them? Who will take temporary responsibility until a new PI takes over?

A2. Responsibilities will be conferred to (in order) 1 – PI, 2 – Co-PI and 3 – one of the two postdoctoral researchers on the project.

Guidance: Indicate a succession strategy in the event that one or more people responsible for the data leaves (e.g., a graduate student leaving after graduation). Describe the process to be followed in the event that the Principal Investigator leaves the project. In some instances, a co-investigator or the department or division overseeing this research will assume responsibility.

Q3. List all expected resources for data management required to complete your project. What hardware, software and human resources will you need? What is your estimated budget?

A3. The project is funded by the Alfred P. Sloan Foundation and received a total of $US 199,977 over a 2-year period. The project received an additional 12 months no-cost extension until March 2023. All costs are in USD and include institutional overhead and benefits. Using TU Delft's Data Management Costing Tool,

we estimate that our RDM costs amount to 10% FTE ($9,360). Together with IT, we therefore budgeted $11,060 for data management.

Guidance: Budgeting. Common purchases are hard drives, cloud storage or software access. TU Delft's Data Management Costing Tool is helpful to determine the share of human labor (FTE) that should be allocated for research data management.

DATA COLLECTION

Q4. What types of data will you collect, create, link to, acquire and/or record?

A4. We will be conducting both quantitative and qualitative research, by way of statistical analysis of bibliographic metadata (i.e., metadata of research datasets and scholarly publications), surveys and semi-structured interviews. The project will collect the following types of data:

Structured numerical and categorical data:

- ➔ Results from bibliometric analysis of data citation and reuse: The data will be collected from DataCite and Dryad. They are a combination of numerical and categorical data and usually available in a structured format, such as XML, JSON or CSV formats.
- ➔ Survey results from researchers: Questionnaires were sent to researchers using their institutional emails sampled from Web of Science corresponding author information. Responses were collected using SurveyMonkey. Results will be analyzed using the statistical analysis software SPSS. The collected data and results will be made available in CSV format after anonymization.

Textual data:

- ➔ Transcriptions from semi-structured interviews: A select number of researchers will be interviewed about sharing, reusing and citing research data. Interviews will be conducted online and recorded and then anonymized and transcribed. Transcriptions will be available in TXT and PDF format. Recordings will be deleted after transcriptions are complete.
- ➔ Codebooks: Interview transcripts as well as free-text answers from the survey will be analyzed and coded using NVivo. The codebook will be available in CSV format.

Audio files:

- ➔ Interview recordings: Interviews will be recorded as audio files in Mp3 format. Audio recordings will be deleted after interviews have been transcribed.

Software and code:

- ➔ Code for processing and analyzing bibliometric data: The bibliometric analyses, including all software code and data sources, will be recorded and documented in Jupyter notebooks. The code will be made available through a Github repository and shared on Zenodo.

Guidance: Data types. Your research data may include digital resources, software code, audio files, image files, video, numeric, text, tabular data, modeling data, spatial data, instrumentation data.

Q5. File formats

A. What file formats do you expect to collect (e.g. .doc, .csv, .jpg, .mov)?
B. Are these file formats easy to share with other researchers from different disciplines?
C. In the event that one of your chosen file formats becomes obsolete (or is no longer supported) how will you ensure access to the research data?
D. Does your data need to be copied to a new media or cloud platform, or converted to a different file format when you store or publish your datasets?

A5 – A: Our data will be collected in the following file formats:

- ➔ CSV (all numerical and categorical data, survey data collected via SurveyMonkey)
- ➔ PDF and TXT (interview transcripts)
- ➔ Jupyter notebooks (.ipynb, the notebook document from bibliometric analysis)

- ➔ Recording files (e.g. MP3, MOV)
- ➔ NVivo data format (coded interview transcripts e.g. .nvpx)

A5 – B: We will make sure that all published data will be available in non-proprietary formats, such as CSV, that are easily shared with and reused by other researchers.

A5 – C: All published data will be available in non-proprietary or open formats (CSV, txt.,.ipynb), which we do not expect to become obsolete without ensuring compatibility.

A5 – D: Copying or converting file formats is not necessary as a final step for data deposit, as all (anonymized) data will be published in accordance with an open science workflow.

Guidance: Proprietary file formats requiring specialized software or hardware are not recommended, but may be necessary for certain data collection or analysis methods. Using open file formats or industry-standard formats (e.g. those widely used by a given community) is preferred whenever possible. Read more about file formats: Library and Archives Canada, UBC Library or UK Data.

Q6. Conventions

A. How will you structure, name and version-control your files to help someone outside your research team understand how your data are organized?
B. Describe your ideal workflow for file sharing between research team members step-by-step.
C. What tools or strategies will you use to document your workflow as it evolves during the course of the project?

A6 – A: All data will be organized via pre-determined file structure, file naming convention and version control. Each project will have a ReadMe file which will document data cleaning and analysis procedures at a high level. Each component (e.g. sampling, data cleaning, or transcripts) will also have a detailed log (TXT format) to document detailed analysis decisions and workflows. We will follow a general agreed-upon file naming convention, as shown below:

- ➔ Raw data (data collected from an instrument, e.g., survey results, interview recording, bibliometric data): stored on Dropbox or a secure server (Microsoft OneDrive) at the University of Ottawa (PI).
  File naming: Content_Raw_YYMMDD.csv (e.g., surveydata_anon_YYMMDD.csv)
- ➔ Processed data (raw data that has been processed, e.g., anonymized interview transcript): stored on Dropbox.
  File naming: Content_Processed_ YYMMDD.csv (e.g., surveydata_anon_relabelled_YYMMDD.csv)
- ➔ Cleaned data (processed data that has been corrected, enriched and analyzed, e.g., anonymized survey results with incomplete cases removed): our working documents will be stored in a shared folder (i.e., Dropbox) to allow all team members to access them. There might be multiple versions of cleaned data of these working documents, but we will use built-in version control of shared folders (i.e., Dropbox) for minor changes. For every major data cleaning step, we will create a new file version and label it using the date to indicate the year (e.g., YYMMDD).
  File naming: Content_Cleaned_YYMMDD.csv (e.g., surveydata_anon_Cleaned_YYMMDD.csv)
- ➔ Final data (cleaned data with documentation): The final data is the final version of the cleaned data plus documentation, such as the ReadMe file, to ensure reuse by others. This version will be published on Zenodo, where it will be archived and downloadable for a minimum of 20 years.
  File naming: Content_Published-v1.csv and Content_ReadMe-v1.txt (e.g., DataCitations_Published-v1.csv and DataCitations_ReadMe-v1.txt)

Guidance: File naming and versioning. It is important to keep track of different copies or versions of files, files held in different formats or locations, and information cross-referenced between files. This process is called 'version control'. Logical file structures, informative naming conventions, and clear indications of file versions, all contribute to better use of your data during and after your research project. These practices will help ensure that you and your research team are using the appropriate version of your data, and minimize confusion regarding copies on different computers and/or on different media. Read more about file naming and version control from UBC Library or the UK Data Service.

A6 – B: We will follow the following general workflow:

1. Data collection
    a. Collect data via instrument (e.g., survey, API)
    b. Document data collection methods in ReadMe/log files. A spreadsheet will be created to manage the survey recruitment rounds and monitor completion statistics.
2. Data storage and preprocessing
    a. Store raw data
    b. If data contains personally identifiable information, store on secure server (Microsoft One Drive), anonymize and deidentify
    c. Document preprocessing procedures in ReadMe/log files
    d. Store anonymized and deidentified data on shared Dropbox folder
    e. Permanent file removal of raw data (e.g., audio recordings) as stipulated in ethics protocol
3. Data cleaning
    a. Clean and prepare data for analysis
    b. Document data cleaning procedures in ReadMe/log files
4. Analysis
    a. Analyze data
    b. Document methods of analysis in ReadMe/log files
5. Publication
    a. Publish data and ReadMe file on Zenodo

A6 – C: When joining the project team, all collaborators will get access to the shared Dropbox folder and be trained on using the DMP via an onboarding document. Onboarding will describe workflows, data storage practices, file naming conventions and RDM responsibilities of each team member.

Guidance: Document workflows. Have you thought about how you will capture, save and share your workflow and project milestones with your team? You can create an onboarding document to ensure that all team members adopt the same workflows or use workflow management tools like OSF or GitHub.

DOCUMENTATION AND METADATA

Q7. What support material and documentation (e.g. ReadMe) will your team members and future researchers need in order to navigate and reuse your data without ambiguity?

A7. Every project will be accompanied by a ReadMe file that documents data collection, processing and cleaning procedures as well as methods of analysis (see Q6). Bibliometric analyses will be documented and carried out using Jupyter Notebooks, which allow for detailed documentation of various steps of data collection (via APIs), processing and analysis in one document. The documentation will be detailed enough that the data and software code can be reused by all team members and, once published, the scientific community and the public.

Every published dataset will be accompanied by a detailed ReadMe file, including a data dictionary that outlines the codes and variables we use in the survey. All qualitative interviews will include a summary including identity of data collector, location of interview and the interview date.

Guidance: Data documentation. Typically, good data documentation includes information about the study, data-level descriptions, and any other contextual information required to make the data usable by other researchers. Other elements you should document, as applicable, include: research methodology used, variable definitions, vocabularies, classification systems, units of measurement, assumptions made, format and file type of the data, a description of the data capture and collection methods, explanation of data coding and analysis performed (including syntax files). View a useful template from Cornell University's "Guide to writing 'ReadMe' style metadata".

Q8. How will you undertake documentation of data collection, processing and analysis, within your workflow to create consistent support material? Who will be responsible for this task?

A8. In order to encourage the reuse of data, methods, results, and to gain early feedback and improve quality, the project adheres to the principles of open science. One of our research goals is to share data and results as early as possible in the research data lifecycle (i.e. DCC Curation Life Cycle model). This will be achieved by publicly documenting the project's progress by publishing all outputs on the project's Zenodo community page and by regularly informing stakeholders. Every team member will be held responsible to continuously document any major processing, updating and analysis of the data in the ReadMe files. Their responsibilities will be highlighted in an onboarding document that each new team member will be expected to read.

While every team member is responsible for documentation as part of the everyday workflow, the lead author or leading researcher on the team who works on data collection and analysis will be responsible to ensure that documentation is sufficient and ReadMe files are complete before the dataset is published. We are using the CRediT taxonomy and follow the lab's authorship guidelines to determine authorship roles and responsibilities at the beginning of each project

Guidance: Establish responsibilities for data management and documentation early on. Some researchers use the CRediT taxonomy to determine authorship roles at the beginning of each project. They can also be used to make a team member responsible for proper data management and documentation.

Q9. Do you plan to use a metadata standard? What specific schema might you use?

A9. Each dataset will be published using the Datacite Metadata Schema 4.3 (or later).

Guidance: Metadata for datasets. DataCite has developed a metadata schema particularly for datasets. It lists a set of core metadata fields and instructions to make datasets easily identifiable and citable. There are many other general and domain-specific metadata standards. Dataset documentation should be provided in one of these standard, machine readable, openly-accessible formats to enable the effective exchange of information between users and systems. These standards are often based on language-independent data formats such as XML, RDF, and JSON. There are many metadata standards based on these formats, including discipline-specific standards. Read more about metadata standards at the UK Digital Curation Centre's Disciplinary Metadata resource.

Q10. How will you make sure that a) your primary data collection methods are documented with transparency and b) your secondary data sources (i.e., data you did not collect yourself) — are easily identified and cited?

A10. We will ensure transparency through documentation in Jupyter Notebooks and ReadMe files. Details are provided in A6-8 and A16.

STORAGE AND BACKUP

Q11. List your anticipated storage needs (e.g., hard drives, cloud storage, shared drives). List how long you intend to use each type and what capacities you may require.

A11. We anticipate using a maximum of 50 gigabytes for all raw, processed, cleaned and final data as well as software code and metadata stored on Dropbox, Microsoft OneDrive Zenodo, and Github.

Q12. What is your anticipated backup and storage schedule? How often will you save your data, in what formats, and where?

A12. All data will be stored on Dropbox or Microsoft OneDrive, which feature version control, in case we need to restore to an older version.

Guidance: Follow the 3-2-1 rule to prevent data loss. It's important to have a regular backup schedule — and to document that process — so that you can review any changes to the data at any point during the project. The risk of losing data due to human error, natural disasters, or other mishaps can be mitigated by following the 3-2-1 backup rule: Have at least three copies of your data; store the copies on two different media; keep one backup copy offsite.

Read more about storage and backup practices from the University of Sheffield Library and the UK Data Service.

Q13. Keeping ethics protocol review requirements in mind, what is your intended storage timeframe for each type of data (raw, processed, clean, final) within your team? Will you also store software code or metadata?

A13. All non-personal data will be stored on a Dropbox shared folder and will be accessible to all team members and collaborators. Files stored on Dropbox are automatically synced between different platforms and users, and are version controlled in case older versions need to be restored. Personal data from interviews and surveys will be stored on secure servers accessible to the PI, Co-PI and team members directly involved in data collection and interview transcription via Microsoft OneDrive. The PI and Co-PI will have permanent access to Microsoft OneDrive, while team members involved in data collection and transcription will only have temporary access to perform anonymization or transcription. If a participant asked to be removed from the sample we did so immediately. Data will be anonymized and de-identified immediately to store anonymized versions of the data in Dropbox, to which all team members have access. All data will be accessible on Dropbox to team members and collaborators until five years after the end of the project, except for the raw data that is permanently deleted as stated above.

DATA SHARING AND REUSE

Q14. Data sharing and reuse

A. How will your data (both raw and cleaned) be made accessible beyond the scope of the project and by researchers outside your team?
B. Is digital preservation a component of your project and do you need to plan for long-term archiving and preservation?

A14 – A: All outputs of the project will be in long-term storage via deposit in data and OA repositories (up to 20 years). We do not anticipate the need for preservation measures beyond this timeframe. All final processed and cleaned datasets as well as select working documents will be published on Zenodo using a CC-BY license to ensure maximum diffusion and reuse. Each Zenodo document will be issued a DOI, which we will use, when citing documents and datasets in related publications.

Raw data will not be published, as most underlying data is already available via open databases and APIs; however, protocols for data collection, processing and cleaning secondary data sources will be documented in important milestone working documents, such as white papers or data papers, published on Zenodo. As mentioned above, raw data for interviews will be permanently deleted.

Zenodo will retain uploaded items for the lifetime of the repository, which equals the lifetime of the host laboratory CERN which is currently financed for a minimum of 20 years. In case of closure of the repository, CERN has committed its best efforts to integrate all content into suitable alternative institutional and/or subject based repositories.

A14 – B: Long term archiving and preservation beyond what is described above is not required for this project.

Q15. What data will you be sharing publicly and in what form (e.g. raw, processed, analyzed, final)?

A15. Most of our raw data is already openly available via publicly-available APIs (e.g., Datacite, Crossref). We will publish all final (de-identified, processed, cleaned and analyzed) data in CSV format on Zenodo, where each dataset receives a DOI to increase findability and reuse. We will publish all software code in Jupyter Notebooks to ensure transparent and open methodology including data collection, processing and analysis.

Q16. Have you considered what type of end-user license to include with your data?

A16. All final data and publications will be published open access, using a CC-BY license wherever possible. By default, our data and publications will be shared using Creative Commons Attribution CC-BY 4.0 International license. If CC-BY is not possible, we will use CC-BY ND.

Guidance: Creative Commons licenses. As the creator of a dataset (or any other academic or creative work) you usually hold its copyright by default. However, copyright prevents other researchers from reusing and building on your work. Creative Commons (CC) licenses are a free, simple, and standardized way to grant copyright permissions and ensure proper attribution (i.e., citation). CC-BY is the most open CC license and allows others to copy, distribute, remix and build on your work, as long as they give you proper credit (i.e., cite your work).

Q17. What tools and strategies will you take to promote your research? How will you let the research community know that your data exists and is ready to be reused?

A17. Our aim is to promote knowledge mobilization as much as possible using traditional publication venues as well as social media. All published datasets and this DMP, as well as presentation slides and preprints, will be aggregated on the project's Zenodo community page so they are accessible from one location. All publications will properly cite underlying datasets by listing their metadata and DOI (according to Datacite Metadata Schema 4.3) in the reference list to allow readers to access and reuse the dataset. The Zenodo community page will be linked to from the project's website at scholcommlab.ca/research/data-citation. We will frequently announce publication of outputs via the PI and Co-PI's social media accounts. The PI's lab, the ScholCommLab, will also announce project outputs and milestones through newsletters and blog posts. Internally, research outputs will be communicated through the ScholCommLab and Make Data Count Slack channels.

Guidance: Knowledge mobilization. Using social media, newsletters, bulletin boards, posters, talks, webinars, discussion boards or discipline-specific forums are good ways to

gain visibility, promote transparency and encourage data discovery and reuse.

Guidance: Help others reuse and cite your data. If you publish your data on a data repository (e.g., Zenodo, Dataverse, Dryad), it can be found and reused by others. Many repositories issue DOIs which make it easier to identify and cite datasets. Did you know that a dataset is a scholarly output that you should list on your CV just like a journal article?

ETHICS AND LEGAL COMPLIANCE

Q18. Are there institutional, governmental or legal policies that you need to comply with in regards to your data standards?

A18. The project will comply with the research data management policies of its host institutions, which take into account relevant legislation, industry standards and best practices. Specifically, we'll be referring primarily to the University of Ottawa's legal and ethical considerations and the Canadian Tri-Council's TCPS 2 (2018), but we may refer to the University of Kiel's integrity and ethics in research policy, if the TCPS 2 doesn't provide enough clarity. As the Co-PI is affiliated with a European institutions, we will comply with the EU's General Data Protection Regulation (GDPR).

Guidance: Compliance. Inform yourself on industry standards and best practices to ensure you are complying with state-of-the-art data management requirements. If you collaborate with a partner in the European Union, you might need to follow the General Data Protection Regulation (GDPR).

Q19. Will you encounter protected or personally-identifiable information in your research? If so, how will you make sure it stays secure and is accessed by approved team members only?

A19. We will store sensitive data on Microsoft OneDrive which is used as a secure server in Canada. Access to the secure location will be limited to the PI and Co-PI for the entire project. Additional team members will be granted temporary access for the time that they are working on data collection and anonymization of sensitive data. (Also see answers to question 13 and 15).

Collection of qualitative and personal data will proceed following formal ethical approval from the uOttawa's research ethics board and will require explicit and informed participant agreement for data sharing following the Recommended Recommended Informed Consent Language for Data Sharing (ICPSR). Social media and other web data will be collected and managed in line with the Association of Internet Researchers' Internet Research: Ethical Guidelines 3.0 document. Datasets will be stored securely with password protection and encryption when assessed as sensitive. Data will be anonymized in reporting, except where explicitly agreed otherwise.

Guidance: Personally-identifiable information. Protecting personally-identifiable information could include strategies like changing your password frequently, storing on a secure server or enabling encryption.

Q20. Before publishing or otherwise sharing a dataset are you required to obscure identifiable data (name, gender, date of birth, etc.), in accordance with your jurisdiction's laws, or your ethics protocol? Are there any time restrictions for when data can be publicly accessible?

A20. The first phase of the study is comprised of publicly available data and will not have any privacy concerns. The second phase includes an online questionnaire and semi-structured interviews with researchers. Survey data and interview transcripts will be stored in a uOttawa Microsoft OneDrive folder, which will be stored on a secure server in Canada and will only be accessible to the PI and Co-PI. Temporary access will be provided to select collaborators who will help in transcribing the interviews and anonymizing the data. Anonymized and de-identified versions of the data will be saved in the team's Dropbox folder for further analysis and publication.

Guidance: Privacy protection. Your institution should be able to provide guidance with local storage solutions. Seek out RDM support at your Library or your Advanced Research Computing department.

Third-party commercial file sharing services (such as Google Drive and Dropbox) facilitate file exchange, but they are not necessarily permanent or secure, and servers are often located outside Canada. This may contravene ethics protocol requirements or other institutional policies.

An ideal solution is one that facilitates cooperation and ensures data security, yet is able to be adopted by users with minimal training. Transmitting data between locations or within research teams can be challenging for data management infrastructure. Relying on email for data transfer is not a robust or secure solution.