PUBLISHING REPRODUCIBLE RESEARCH OUTPUTS

A Data Management Plan created using DMPonline

Creators: Lucia Loffreda, Andrea Chiarelli

Affiliation: Research Consulting

Funder: Knowledge Exchange

Abstract: The aim of this project is to investigate current practices and barriers related to publishing reproducible research outputs and to determine how infrastructure (technical and social) can support the vision of an open research enterprise. As part of this work, we will conduct a literature review, focus groups and interviews focusing on key issues faced by researchers, funders, and infrastructure providers in relation to publishing reproducible research outputs.

Template: DCC Template

ORCID iD: https://orcid.org/0000-0002-3548-3124; https://orcid.org/0000-0001-7336-8330

Last modified: 16-04-2021

PUBLISHING REPRODUCIBLE RESEARCH OUTPUTS

DATA COLLECTION

What data will you collect or create?

This research project involves primary data collection via: a literature review of academic and grey literature sources, semi-structured interviews and focus groups and Twitter text and data mining (TDM).

- We will collect publicly available information during the desk-based literature review. Sources are expected to include: journal articles and reports, blog and other social media posts, as well as position statements with a specific focus on publishing reproducible research. Sources will be captured using Zotero, including literature metadata.

- Interviews and focus groups carried out during the project will be recorded and transcribed using Business Friend transcription services. Video files and transcriptions will be stored as on Research Consulting's OneDrive for Business account. Video files will be deleted after 24 months after the end of the project, as per Research Consulting's Data Protection and Privacy Policy.

- Twitter TDM will be carried out via R Studio and documented in an R Notebook. This will include the creation of a csv dataset of tweets.

Based on the data we collect during this project, we expect to create:

- Coding summaries in NVivo, our qualitative coding software, including:
    - Coding summary of literature gathered during the literature review
    - Coding summary of transcripts from focus groups and interviews
- Text and data mining algorithms (R Notebook) and data exports in csv format (in line with Twitter's terms of use)

- [Optional] A poster and/or an academic article, including underlying information or research data

How will the data be collected or created?

Data will be collected via the following methodologies during this research project: structured literature review, semi-structured interviews and focus groups, and qualitative coding using NVivo.

- The structured literature review will consider academic and non-academic articles, blog posts and other sources relevant to our research topic. Sources will be collected using Zotero.

- Recordings will be created from interviews and focus groups using Zoom. Transcriptions will be commissioned and then stored in Word format.

- NVivo will be used to conduct qualitative coding. The qualitative coding created during this project will be stored as NVivo project files.

- TDM will be carried out using Twitters API and the rtweet package.

In all cases, we will use file versioning to track changes to the data. This includes:

- Including a version number, e.g. "v1" or "v2"

- Including the initials of the team member making changes to the file

The consistency and quality of data collection will be controlled and documented via the following processes:

- Team members responsible for the literature review will test the literature section approach for consistency.

- All data collected during the course of the project will be reviewed by the project lead and project director prior to analysis.

- All data collected during the course of this project will be reviewed by the Knowledge Exchange team prior to sharing.

DOCUMENTATION AND METADATA

What documentation and metadata will accompany the data?

To inform the literature review, a detailed literature search strategy will be prepared including information on: search strings used, a taxonomy of subject areas of focus, and a description of selection criteria for sources considered during the review.

Any TDM code used will be shared including documentation as a commented R notebook.

Metadata will be prepared following the schemas implemented by the chosen data repository solution, including cross-linkages between different project outputs using metadata fields. Zenodo's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms. More information on DataCite's Metadata Schema is available via the following CESSDA documentation:
https://www.cessda.eu/content/download/834/7776/file/CMM_ServiceProvidersMetadataPractices_2016

ETHICS AND LEGAL COMPLIANCE

How will you manage any ethical issues?

Our study will involve human participants and we will capture their views on publishing reproducible research outputs via semi-structured interviews and focus groups. Personal information such as names and email addresses will be

collected during the project, alongside affiliation and, in some cases, subject areas. However, we will share our project data in anonymised form, and the list of project participants will only be available in aggregate form (participant names, affiliations and subject areas).

Before participants contribute to our study, they will be fully informed about how their data will be managed via briefing documents shared via email. Participants will be made aware that their contributions via interviews and focus groups will undergo thematic analysis. They will also be made aware that their contributions will be shared in the forms noted in other sections of this DMP (particularly, coding summaries) and that verbatim quotes may be reflected in our project outputs. At the design stage, it is not possible to determine whether attribution may be beneficial or necessary. Therefore, should we wish to include attributable quotes in our outputs, we will seek explicit consent from individual participants and provide them with a chance to review and edit the text.

- We will obtain verbal consent from participants at the start of interviews and focus groups. Interviews will be recorded and transcribed, and verbal consent will therefore be captured in the recording and transcript. Transcripts will be stored in Word format on Research Consulting's OneDrive for business account with only project team members having access to these files.

We are aware of the potential to breach of data protection/confidentiality protocols. In order to address this, we have a GDPR-compliant Data Protection and Privacy Policy and a number of processes and workflows to ensure we gather and process personal information in a lawful way. This covers our entire contacts list, our project management system, the software we use (e.g. NVivo) and our working folders. We are also Cyber Essentials certified.

This study has obtained ethical approval from the University of Oxford via CUREC.

All research data (bar quotes, which are discussed above) will be fully de-identified prior to sharing in the public domain; as per the guidance from the Information Commissioner's Office (ICO), "where an organisation converts personal data into an anonymised form and discloses it, this will not amount to a disclosure of personal data. This is the case even though the organisation disclosing the data still holds the other data that would allow re-identification to take place." The following specific considerations apply:

- Quotes coded via NVivo will not be attributed to individuals and any identifiable information will be removed from the quotes. This may include, for example, affiliations and other information that might allow identification in combination with other data shared in the context of this project (e.g. the list of project contributors). No video recordings nor full transcripts will be shared.

- The data underpinning text and data mining will be in the form of tweet identifiers and is therefore anonymous by default. To reproduce our results, users will be required to carry out tweet hydration, as Twitter's Developer Agreement and Policy does not allow the sharing of full datasets obtained via its API:

    - "If you provide Twitter Content to third parties, including downloadable datasets or via an API, you may only distribute Tweet IDs, Direct Message IDs, and/or User IDs."

    - "In total, you may not distribute more than 1,500,000 Tweet IDs"

How will you manage copyright and Intellectual Property Rights (IPR) issues?

We expect no issues relating to Intellectual Property Rights resulting from the consultation run in the context of this project. Knowledge Exchange will hold the copyright for all outputs produced throughout the project and all outputs shared will use CC-BY licences.

Our literature review will include some sources where IPR are a concern, e.g. subscription or copyrighted articles. Qualitative coding of these will not be shared if the word count for the source under consideration exceeds 400 words: although this is an arbitrary threshold, as noted by the UK copyright service, it was used as a general rule of thumb. We note that the purpose of this project is to benefit the research community and not to derive any financial or other benefits from the sharing of any text extracts.

Examples: if Article A is a subscription or copyrighted article and we have coded <400 words, then the coding extract will be shared; if Article B is a subscription or copyrighted article and we have coded >400 words, then no coding extracts will be shared. This approach is in line with fair use

STORAGE AND BACKUP

How will the data be stored and backed up during the research?

All project data will be stored on Microsoft OneDrive cloud servers, which are a part of the Microsoft Office 365 package. These servers are located in the territory of the European Union and subjected to GDPR.

We have a robust file security system and internal data is securely stored in the cloud (OneDrive for Business), with local encrypted backups also being maintained. We are also Cyber Essentials certified.

How will you manage access and security?

All types of data collected and created during this project will be protected through our secure systems and processes.

Access to un-anonymised research data will be limited to the core project team (outlined in the Responsibilities and Resources section of this data management plan).

SELECTION AND PRESERVATION

Which data are of long-term value and should be retained, shared, and/or preserved?

Personal data collected during this projected will be deleted two years after the project's completion as per Research Consulting's Data Protection and Privacy Policy. This includes video recordings, interview and focus group transcripts, un-anonymised NVivo files and any tweets harvested.

Data of long-term value which will be retained after the end of this project will include:

- Literature review search strategy including search strings, taxonomy of subject areas, and a description of selection criteria
- Literature review database in csv format (exported form Zotero)
- Anonymised coding summaries from NVivo
- R Notebook and underlying data (in line with Twitter's terms of use)

There is also the potential for further data of long-term value to be preserved after the end of this project. Such data could include:

- Material prepared for conference presentations for example slides or poster presentations
- Academic articles

We expect to share the above data publicly, but always in anonymised form and with no attribution.

What is the long-term preservation plan for the dataset?

Data will be shared publicly throughout the course of this project. We will share data as the project evolves, as each phase is completed. Sharing of outputs will take place via Zenodo and all outputs will be linked via metadata fields under the same Zenodo community: https://zenodo.org/communities/ke-prro. ORCID identifiers will be used wherever possible.

Project outputs will also be shared on the Jisc online repository.

Both Zenodo and the Jisc repository have digital preservation systems in place. Zenodo is a FAIR repository solution and complies with the Plan S requirements for open access repositories.

DATA SHARING

How will you share the data?

Data will be shared publicly throughout the course of this project. We will share data as the project evolves, as each phase is completed. Sharing of outputs will take place via Zenodo and all outputs will be linked via metadata fields under the same Zenodo community: https://zenodo.org/communities/ke-prro. ORCID identifiers will be used wherever possible.

Project outputs will also be shared on the Jisc online repository.

All data will be shared in anonymised form and will therefore not affect the privacy of participants. This applies to coded interview and focus group transcripts in particular.

All text coded as part of our literature review will be shared in the form of coding summaries (csv literature export), with the exception of subscription-only or copyrighted materials where a significant portion of text has been coded.

Are any restrictions on data sharing required?

No restrictions to data sharing are required to deliver this project and all data will be anonymised.

RESPONSIBILITIES AND RESOURCES

Who will be responsible for data management?

The core project team will be responsible for data management and implementing this data management plan. These core project team comprises:

Individual	Project role	Key responsibilities
Rob Johnson	Project director	Quality assurance
Andrea Chiarelli	Project lead	Implementation of the present DMP, compliance with relevant regulation, overall supervision of the research team
Lucia Loffreda	Researcher	Data gathering, processing, analysis and reporting
Laura Fortunato	Critical friend	Critical review of project outputs and provision of expert advice

What resources will you require to deliver your plan?

We will not require additional hardware or software to deliver this plan.