European Research Council
Executive Agency

Established by the European Commission

European Research Council (ERC)

ERC Data Management Plan

Template

European Research Council
Established by the European Commission

ERC OPEN RESEARCH
DATA MANAGEMENT PLAN (DMP)

Project Acronym	Project Number
SoundAct	101053194

Template for the ERC Open Research Data Management Plan (DMP). The following sections should describe how you plan to make the project data Findable, Accessible, Interoperable and Reusable (FAIR). Each of the following five issues should be addressed with a level of detail appropriate to the project.

SUMMARY *(dataset[1] reference and name; origin and expected size of the data generated/collected; data types and formats)*

The dataset name is 'ERC Grant 101053194: SoundAct - The actuation of sound change' (henceforth the *SoundAct* dataset when referring to the entire data generated over the lifetime of SoundAct). The *SoundAct* dataset reference is established with this permanent link to identify the project: https://epub.ub.uni-muenchen.de/view/eu/erc-advanced-101053194.html within Open Access LMU at the host institute. This link lists all the publications associated within the *SoundAct* dataset. Each publication will list will be one DOI for a dataset including primary and secondary data (see 1. and 2. below for more details) that were used for the analysis in the publication.

The *SoundAct* dataset originates from spoken language recordings made in Greece, Italy, Japan, and Kenya (Table 2, pdf page 98, grant agreement).

The total expected size of the *SoundAct* dataset that includes 640 targeted speakers (80 speakers × 2 dialects × 4 languages: pdf page 100, grant agreement) and 100 MB per speaker is between 50-75 GB.

Each primary dataset contains the following types of data: .wav sound files, time-aligned derived acoustic and/or physiological files, annotated data with time-stamps structured using the Emu Speech Database Management System: https://ips-lmu.github.io/EMU.html and https://doi.org/10.1016/j.csl.2017.01.002 and details of the anonymized social-indexical information of each recorded subject stored either as a plain text or excel file. The secondary dataset that is derived from the primary dataset is accessible within the R programming environment together with R-readable .Rmd and .html files of which an example is here: https://data.ub.uni-muenchen.de/363/. The .Rmd and .html files of the secondary dataset also contain the commands for accessing the data that was analysed in the publication from the corresponding primary dataset.

[1] *Several datasets may be included into a single DMP.*

1. MAKING DATA FINDABLE *(dataset description: metadata, persistent and unique identifiers e.g., DOI)*

1.1. As stated under the summary above, the *SoundAct* dataset is findable via this permanent link https://epub.ub.uni-muenchen.de/view/eu/erc-advanced-101053194.html. Each publication contained within that link will list one DOI for both the primary and secondary datasets that was used within that publication.

1.2. Each primary and secondary dataset associated with a publication will be given a DOI within an institutional repository provided by the University library of LMU Munich: see Open Data LMU https://data.ub.uni-muenchen.de and Discover https://discover.ub.uni-muenchen.de/

1.3 All publications - and therefore the DOI of the primary and secondary datasets associated with each of *SoundAct*'s publications - are also accessible via the project website here: https://www.phonetik.uni-muenchen.de/Forschung/soundact/soundAct.html#publications

2. MAKING DATA OPENLY ACCESSIBLE *(which data will be made openly available and if some datasets remain closed, the reasons for not giving access; where the data and associated metadata, documentation and code are deposited (repository?); how the data can be accessed (are relevant software tools/methods provided?)*

2.1 The primary datasets are not openly available. This is because (1) the *primary* datasets contain audio files that could allow a person's identity to be established and (2) the *primary* dataset contains information about the social characteristics of the speaker as described under 'Participants' on pdf page 100 of the grant agreement. Further reasons for the need to preserve anonymity were discussed on pdf page 17, point 7 of PROPOSAL_101053194-SoundAct-ERC-2021-ADG.pdf under 'Compliance with ethical principles and relevant legislations'.

2.2 The data, medadata, documentation and code for each primary and secondary dataset are deposited at Open Data LMU https://data.ub.uni-muenchen.de.

2.3 The secondary datasets are accessible from Open Data LMU https://data.ub.uni-muenchen.de

3. MAKING DATA INTEROPERABLE *(which standard or field-specific data and metadata vocabularies and methods will be used)*

3.1 The registration of the DOI is accomplished in co-operation with the University Library, LMU Munich using the DataCite Metadata Generator (https://dhvlab.gwi.uni-muenchen.de/datacite-generator/).

3.2. Viewing, listening to, further processing and/or tagging any *primary* dataset are inter-operable tasks because they require only a web-browser and the R programming language (which are accessible from all major platforms and operating systems for free) but no other stand-alone software. This is because each *primary* dataset is structured using the Emu Speech Database Management System that has been developed at the host-institute since 2006 and which is accessible using a web application with direct interface to the R programming environment, as explained here: https://ips-lmu.github.io/The-EMU-SDMS-Manual/chap-emu-webApp.html.

3.3. Each *secondary* dataset contains commands for replicating all the analyses and figures in each publication. This replication is interoperable both because it makes use of the R programming environment and because all the commands for doing so are stored as R-derived .html documents that can therefore be accessed from any web browser.

4. INCREASE DATA RE-USE *(what data will remain re-usable and for how long, is embargo foreseen; how the data is licensed; data quality assurance procedures)*

4.1 *Re-use*. The data stored on the LMU University Library servers is made available through APIs for access through BASE Academic Search, Google Dataset Search, and other similar services.

The re-usability of the *primary* datasets is guaranteed as long as web-browsers like Chrome and the R programming environment continue to exist. The re-usability of the audio data in the *primary* datasets is highly likely in the future since these are stored as .wav files which is an audio file format standard that has existed since 1991 and which is recommended as a file type for long-term storage (e.g., https://documentation.library.ethz.ch/display/DD/Archivtaugliche+Dateiformate).

The re-usability of the *secondary* datasets depends on the continued existence of the R programming environment and the HTML markup language. The re-usability of the secondary datasets does, however, assume that the R packages that were used for their analysis continue to exist. There are facilities within the R programming environment for the installation of older packages e.g., https://search.r-project.org/CRAN/refmans/remotes/html/install_version.html. In addition, the *renv* package https://rstudio.github.io/renv/articles/renv.html can be used to re-create older R environments. Taking into account these points, the *primary* and *secondary* datasets are highly likely to be re-usable for at least a decade beyond the end of the project.

4.2. *Embargo*. Access to the *primary* dataset will be restricted under a conditional embargo to protect individual identity. 'Conditional' means here that the datasets will be made openly available for registered academics after 10 years as soon as the owner is not able to manage the user access to the primary datasets any longer. This policy is a common practice in scientific data repositories to avoid archived datasets that are never re-used in the future (which would contradict the FAIR principles).
The secondary datasets do not require an embargo since they are open accessible from the start.

4.3. *Data quality assurance.* The audio recordings will take place in quiet environments, using state-of-the-art recording equipment and a high sampling rate (in general, 44.1 kHz with 16 bit linear quantization, mono or stereo) to ensure the high audio quality needed for phonetic research. The file format chosen for the *secondary* datasets, namely R Markdown, is specifically designed to write clean code with sufficient explanatory text so that the analyses can be easily replicated later on.

5. ALLOCATION OF RESOURCES and DATA SECURITY *(estimated costs for making the project data open access and potential value of long-term data preservation; procedures for data backup and recovery; transfer of sensitive data and secure storage in repositories for long term preservation and curation)*

5.1 *Estimated costs for making the project data open access*. There are no costs involved in the storage of the datasets at the University Library of LMU. There are, of course, also costs associated with making publications open access. These publication costs have been budgeted for within SoundAct as detailed under 4.1, pdf page 118 of the grant agreement.

5.2 *Potential value of long-term data preservation*. Speech data preserved over the long-term is increasingly used to explain the mechanisms by which the use of speech and language changes over time.

5.3 *Procedures for data backup and recovery*. The University Library of LMU that is the repository of the datasets has daily backups via the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities https://www.lrz.de/english/.

5.4 *Secure storage in repositories for long term preservation and curation*. There are well-established procedures in place at the University Library of LMU, and the LRZ as explained here https://doku.lrz.de/backup-und-archivierung-10745679.html for the long-term preservation of the datasets. In addition, a data scientist officer who is in a permanent position at the Institute of Phonetics and Speech Processing has been named and has agreed to be responsible for the long-term curation of the data.

DISCLAIMER. Please note that the ERC Data Management Plan is not a part of the Ethics Review. It is the responsibility of the Principal Investigator to inform the ERCEA Ethics Team of any ethics issues/concerns regarding the collection, processing, sharing and storage of data in relation to the project.