AGENCE NATIONALE DE LA RECHERCHE (ANR): ANR - DMP PROJECT: EMBER

Main author: Pierre Dragicevic (Inria Saclay Île-de-France, team Aviz)
Contributed: Yvonne Jansen (CNRS, Sorbonne Université), Petra Isenberg (Inria Saclay Île-de-France, team Aviz), and Laurent Romary (Inria Paris, team ALMAnaCH), with the help of the Inria library network staff.
Date: 07.10.2020
Version: 1.1

This data management plan is for the EMBER project funded by the 2019 ANR PRC call (ANR-19-CE33-0012, ember.inria.fr/). The project is a collaboration between Inria Saclay, Inria Bordeaux, and Sorbonne Université.

This data management plan is based on the template provided by ANR:
dmp.opidor.fr/template_export/1858712127.pdf

1. DATA DESCRIPTION AND COLLECTION OR RE-USE OF EXISTING DATA

1a. How will new data be collected or produced and/or how will existing data be re-used?

Data will be collected through user studies, which will be the primary means by which we will understand user needs, evaluate the extent to which the prototypes we develop address these needs, and answer general research questions related to the project.

We expect the whole project to involve about 10–20 user studies conducted by different teams (PhD students, interns, post-docs, their advisors, and potential external collaborators). Each study will address research questions identified based on the team's interests, the current state of the art, and the findings of prior EMBER studies. Therefore, it is impossible to decide the exact nature of these studies ahead of time.

As stated in the project proposal, we will draw from a broad range of inquiry methods commonly used in human-computer interaction (HCI) and visualization research, including interviews, surveys, overt observation and feedback sessions, and controlled experiments. The studies will take place either in the lab, in public spaces (e.g., museums) or online (e.g., on crowdsourcing platforms). The type of data collected will be qualitative and/or quantitative depending on the study.

Although we may reuse existing data from previous studies that was made available to researchers (e.g., on open science repositories), our focus will be on collecting new data.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

The data will depend on the nature of each study. For example, online behavioral studies and user interviews typically produce very different kinds of data. Overall, across all types of user studies we are able to anticipate, we will collect as raw data:

- Data tables of behavioral measurements (e.g., reaction times, mouse clicks, …) and responses to question items collected in controlled experiments (in tabular format such as CSV or json)
- Video recordings of interviews and overt observational studies (mp4, mkv, avi, ...)
- Audio recordings of interviews and overt observational studies (mp3, ogg, ...)

This raw data will then be processed into:

- Other data tables that will contain derived measurements, human-generated codes, aggregated measurements, and statistics (in tabular format such as json or CSV)
- Data summaries in human-readable form such as plots and numerical tables (e.g., PDF, R markdown,...)
- Interview transcripts (.txt)

2. DOCUMENTATION AND DATA QUALITY

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

There is no metadata standard in our research field, but we will document:

- the provenance of all our data files, as well as the data collection methodology
- how to interpret our data files, including the meaning of each column in data tables
- the data processing and analysis procedures (whether human or algorithmic)
- the directory structure used to store data and code

The documentation will be written in readme files and lab notebooks (e.g., R markdown, Jupyter notebooks), and stored and shared with the data (see Sections 3 and 5).

2b. What data quality control measures will be used?

Although there is no data quality control standard in our field, in all our studies we will strive to maximize the quality and reliability of our data through rigorous study design, extensive pilot testing, and whenever possible, power analysis and pre-registration of experiment protocols and analyses. The particular methods will be decided on a per-study basis, and documented in the academic publication where the study is reported.

3. STORAGE AND BACKUP DURING THE RESEARCH PROCESS

3a. How will data and metadata be stored and backed up during the research?

During the course of each study, small-size anonymous, anonymized and pseudonymized data files will be stored on a private repository on gitlab.inria.fr, together with the rest of the research material (data analysis code, code for the experimental software) and the documentation.

Large media files such as videos and audio interviews will be stored on mybox.inria.fr (hosted by Inria) or mycore.core-cloud.net/ (hosted by CNRS).

Data related to identified or identifiable persons as well as all information allowing to re-identify pseudonymized data will be stored in password-protected encrypted directories on mybox.inria.fr or mycore.core-cloud.net/ (one per study). Only the investigators directly involved in the study and listed in the institutional ethics board application will be granted access.

All these file hosting and versioning services have automatic backup mechanisms.

For data collected in the form of physical artifacts (e.g., filled questionnaires, artefacts created by participants), a digital scan will be made and stored as described above, while the original artefacts will be stored in a locked cabinet accessible only to the investigators.

3b. How will data security and protection of sensitive data be taken care during the research

Non-sensitive anonymous data generated as part of each study will be made accessible to all collaborators involved in the study, while the research is ongoing (for after the research is completed, see section 5).

As mentioned above, personally identifiable data will be stored in an encrypted institutional repository or a locked cabinet to which only the investigators explicitly listed in the institutional ethics board application will be granted access. Investigators will be instructed not to keep digital files in local storage devices for longer than the time necessary to create them or process them (after which they will need to be promptly and securely deleted).

4. LEGAL AND ETHICAL REQUIREMENTS, CODE OF CONDUCT

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

All studies involving human participants will be submitted for approval to an institutional ethics board before the study is conducted. The available committees in the EMBER consortium are the COERLE at Inria (Comité opérationnel d'évaluation des risques légaux et éthiques), the CER Paris-Saclay (Comité d'éthique pour la recherche de l'Université Paris-Saclay) and the CER Sorbonne Université (Comité d'éthique de la recherche de Sorbonne Université). The choice of ethics board will be done on a per-study basis, depending on the affiliation of the involved investigators.

We will pay special attention to data protection issues that may arise in the course of the project. In particular, if a study requires the collection of personally identifiable data, we will work in close collaboration with our institutional lawyers and data protection officers ahead of time to ensure that the GDPR will be respected.

All studies will require informed consent as mandated by ethics boards. Studies involving the collection and storage of personal data will use GDPR-compliant consent forms (e.g., as produced by https://consent.dariah.eu). For in-person studies, signed consent forms will be collected and stored as described in section 3a, for a period of three years.

Since the goal of the EMBER project is to study situated visualizations of personal data, some of our studies will involve giving participants tools that allow them to visualize their own data (e.g., electricity consumption, fitness data, e-mail activity,...). In such studies, participants' own data will never be collected by us. We will only collect data about the way participants use our prototypes and their feedback about our prototypes. The type of personally identifiable data we may collect includes, for example, video interviews.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

All the data generated during the course of the EMBER project will be empirical data meant to advance scientific knowledge. Thus we will not seek any intellectual property protection for the data generated, but instead will adhere to the principles of open scientific inquiry, requiring that we make our data freely available to other researchers for the purposes of scrutiny, re-analysis and replication. Data will be released publicly whenever possible (see Section 5), and when this is not possible, it will be destroyed after a predefined time period as per GDPR regulations.

For each study, a principal investigator will be identified who will be in charge of the data collected during the course of that study.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

We will ensure that all human subjects who produce data for the EMBER project are protected according to commonly accepted international standards of ethics, which require among other things informed consent and a fair assessment of risks and benefits (*National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*). This will be enforced by submitting the protocol of all our studies to an institutional ethics board. No study involving human subjects will start without approval from an ethics board.

5. DATA SHARING AND LONG-TERM PRESERVATION

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Each study involving the collection of data from human subjects will be eventually published as a peer-reviewed article. When the article is first submitted for peer review, anonymized data files and other research material relevant for peer review will be uploaded as a private project on the German servers of OSF (the Open Science Framework, hosted at https://osf.io/, which is the repository of the Center for Open Science). A link to an anonymized view of the project will be included in the submitted article. Once the article is accepted, the OSF project will be made public. The data will remain on OSF for as long as technically possible (in the worst case, OSF has a 50+ year data preservation fund). Since all the data is anonymized, it is outside the scope of the GDPR, and the right to erasure and retention periods do not apply.

All data related to identified or identifiable persons as well as all information allowing to re-identify pseudonymized data will be kept for a predefined period of time after publication and then permanently erased (or physically destroyed, for artefacts such as paper). The duration will be decided on a case by case basis: it could be for example one or two years to give the PhD student time to defend. Signed consent forms will be kept for three years after the study.

Participants will have the possibility to access and erase their personal data at any time during and after the study. Such events will be explicitly documented (e.g., in the publications and in the PhD thesis) should they interfere with the research plan.

The project will adhere to Inria’s and Sorbonne Université's policy of open access requiring that all publications are made available in the HAL repository (hal.archives-ouvertes.fr/).

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

As mentioned previously, (i) anonymized data will be stored on the servers of the Center for Open Science (COS, https://osf.io/) ; COS has a preservation fund which will preserve and maintain read access to hosted data for 50+ years should COS shut down. (ii) Data related to identified or identifiable persons will never be shared outside the authorized investigators and will eventually be permanently erased.

5c. What methods or software tools are needed to access and use data?

Upon publication of each study, anyone (researchers or members of the general public) will be able to browse or download the anonymized data and other research material from the OSF website. Files will be in a standard format and will be documented (see Sections 1 and 2).

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

We will use URLs to OSF projects as a primary means of linking to our research material, as this will allow us to update our content and possibly correct errors. Whenever we need a persistent identifier pointing to a fixed version, we will use DOIs, which are supported by OSF.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

The responsibility concerning the data collected for EMBER will lie with the PI of the project (Pierre Dragicevic). The PI will also make sure that this DMP is implemented and updated. In addition, for each study, a principal investigator will be identified who will be in charge of the data collected during that study and who will report to the PI. This principal investigator will need to be a permanent researcher (typically, the PhD advisor).

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

There will be no material cost besides the 17,500€ worth of personal computers that will be funded by EMBER and used to collect, temporarily store, and process the data. The online data storage, versioning and sharing services we will rely on are either funded by French research institutions (gitlab.inria.fr, mybox.inria.fr, mycore.core-cloud.net/) or by international non-profits (osf.io/). In terms of time resources, preparing the data for sharing is an integral part of the projects funded by EMBER and will be handled for each deliverable by the PhD students, postdocs, and their advisors.

After completion of the EMBER project, the data maintenance costs will be split among (i) the Center for Open Science (COS) who funds the service osf.io/ on which all the open material will be shared; and (ii) the partner institutions (Inria, Sorbonne Université) who employ the PIs who will dedicate a small fraction of their time to respond to requests concerning the open material, improve the documentation if needed, and fix possible errors.