Università
della
Svizzera
italiana

Research and Transfer Service

Data Management Plan for SNSF research projects

Examples

12.04.2022, IS, version 2.0

Contents

Introduction

This document offers some examples of Data Management Plans (DMP) submitted to the Swiss National Science Foundation (SNSF) together with the respective research projects. The examples presented are diverse in their nature and form, and represent all the academic disciplines of USI.
This series of examples complement the SNSF DMP guidelines that the user may find at the dedicated USI Research and Transfer Service (SRIT) website.

SNSF DMP examples

Architecture

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

Based on historical documents and images, this qualitative research will produce georeferenced data sheets of a group of disappeared historical buildings. According to the documents, data will include information about the construction site, the building, the architects involved and the collections inside the buildings. Bibliographic research will allow to compare and include data produced by previous research. The data produced from this research will be:

1. bibliographical files;
2. transcriptions of archival documents;
3. historical research records;
4. images, photographes (about 1200);
5. images catalogues;
6. images lists.

Data in categories 1, 2, 3, 5, 6 will be documented in Word, and also then Pdf, file format. Data in category 4 will be documented in Tiff, and also then Pdf, file format. The data produced in category 4 will amount to approximately to 80 GB.

1.2. How will the data be collected, observed or generated?

Following the approach of history of architecture, data sheets will provide descriptions and data organized according to standard systems such as Bildindex (https://www.bildindex.de/). Documents will be organised using the university internal system of file sharing; a selection of documents and images will be made available to a broader public on the project website. The images will be purchased as already digitized by their rights holders.

1.3. What documentation and metadata will you provide with the data?

We will provide all data sheets, bibliographical files, transcriptions of archival documents and images catalogue and list. To allow a broader audience to access the data generated by the project, the research will contribute to Wikidata, the open, free knowledge and structured repository feeding also the WIkimedia projects. Wikidata allows to share content with metadata and it has standard description systems for buildings. Further content will be shared on the project website to contextualise the research, to provide a selection of reproductions of documents and significant images.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

This historical research in the field of architecture does not present challenges related to ethical issues.

2.2. How will data access and security be managed?

This historical research in the field of architecture does not present challenges related to personal or other sensitive data. Data produced by the project can be released under CC0.

2.3. How will you handle copyright and Intellectual Property Rights issues?

Data produced by the project is released under CC0 and on Wikidata. Content produced by the project is in CC BY-SA license and articles are published under gold or green open

access. There might be restrictions in publishing historical documents and images due to the involvement of third parties; the discussion about how to release sources will be addressed during the project and it is possible that documents might be available only for publication on the project's website.

3. Data storage and preservation
3.1. How will your data be stored and backed-up during the research?

The project data will be stored and processed on file servers managed by the IT service of USI, which are protected by backup copies of data (backup) made every night.

3.2. What is your data preservation plan?

The data preservation plan of the research is based on sharing data and content on existing platforms. The project does not create a new database but it uses existing systems; this approach increases the potentiality of reuse of the project's data and it relies on multiple open repositories with extensive communities. Content related to the project will be also shared on an internal university website which is maintained directly by the university as a subdomain.
To guarantee the durability in time, where possible, we will store files also in open archival formats, e.g. Word and Tiff files converted to PDF.

4. Data sharing and reuse
4.1. How and where will the data be shared?

Due to its humanistic nature, data of this research will be shared on repositories specialised in art history or addressing a broader audience. Wikidata – even if it is a peculiar repository and not a typically academic one – complies with the FAIR Data Principles and it is non-commercial. Eventually data can also be shared on Zenodo. The research content will also be published on an internal website managed by the university with texts under CC BY-SA.

4.2. Are there any necessary limitations to protect sensitive data?

This historical research in the field of architecture does not present challenges related to sensitive data. Images under copyright will be available only by agreement with the rights holders. The discussion about how to release images will be addressed during the project and it is possible that these documents might be available only for publication on the project website.

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Biomedicine #1

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

The new data generated will mainly consist of:

- images originated by microscopy analyses (e.g. immunofluorescence, immunohistochemistry, live-imaging acquisitions) in .tiff format. In addition, raw data in a format depending on the specific microscope software will be generated. All the images generated will potentially occupy a space ranging from 100 GB to 1 TB;
- spreadsheets and free-text files, in .xlsx and .txt format. These files will presumably occupy no more than 1 GB;
- files originating from transcriptomic analyses, mass spectrometry and secretome analyses in format .csv. These files will potentially occupy no more than 1 GB.

In addition, during the data correlation step relevant results from literature and biological databases (e.g. GEO, STRING) will be analyzed, organized in concept maps and used to guide, interpret and validate the analyses performed in the project.

1.2. How will the data be collected, observed or generated?

All samples on which data are collected will be prepared according to published standard protocols in their respective field.
Images will be collected using different microscopy techniques (e.g. optical, epifluorescence, confocal). Microscopes used will be regularly serviced and re-calibrated. During image acquisition particular care will be considered for setting uniform imaging parameters and guarantee the possibility to properly compare different images.
Regarding transcriptomic, mass spec and secretome analyses, internal reference standards will be used with appropriate controls. In particular for transcriptomic analyses, special care will be considered to guarantee the highest possible quality of the genetic material.
For each data, at least three experimental replicates will be performed. Technical replicates will be also provided.
Data originated from each experiment will be stored in dedicated folders, numbered in progressive order and sub-structured with specific folders for each data format (e.g. a folder for immunofluorescence images, a folder for transcriptomic data).
Successive versions of the same file will be named as the original file with a progressive number accounting for the number of the version.

1.3. What documentation and metadata will you provide with the data?

Each image, transcriptomic dataset, mass spec dataset or secretome dataset will be enclosed within a file including the name and a persistent identifier for each file, the name of the person who collected or contributed to the data, the date of collection and the conditions to access the data.
Files will be named according to a pre-agreed scheme. The dataset will be accompanied by a README.txt file which will describe the directory hierarchy and the file naming scheme. Each directory will contain an INFO.txt file describing the experimental protocols used in that specific experiment. It will also record any deviations from the protocol and other useful contextual information.
Microscope images store a wide range of metadata (e.g. field size, magnification, lens phase, zoom, gain, pinhole diameter) for each image. This should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future. Regarding omic analyses, metadata stored by the specific equipment employed (e.g. next generation sequencer, mass spectrometer) will be

complemented with detailed description of the samples that were analyzed (e.g. source of the biological tissue, specific cell line, number of replicates).
The researchers who will be in charge of each specific experiment and its correlated analyses will be responsible for checking that metadata are properly stored and accessible.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

During the project I will use commercially available human plasma samples and cell lines (e.g. endothelial cells, fibroblasts). These samples are collected by specific companies under informed consent.
During the validation step, I will perform image analyses on human tissue samples (e.g. tissue microarrays) which are commercially available. These samples are collected under informed consent.
Overall, only limited anonymized data including sex, age, body mass index and presence of diabetes or cardiovascular diseases will be collected from the supplier.

2.2. How will data access and security be managed?

All the data of the project will be stored on institutional centralized IT facilities, which thus follow the Swiss standards required for hospitals.
Only experimental data obtained using anonymized data will be used and stored.
The data will be stored on dedicated folders under the IT infrastructure, with limited access only to laboratory personnel. Daily backups will be performed by IT service.

2.3. How will you handle copyright and Intellectual Property Rights issues?

Intellectual property of the data belongs to the Institution in which the data will be generated (i.e. USI). In case I will foresee a possible exploitation of the generated data, an IPR agreement will be established between parties.
In case third-party data are used during the correlation analyses or the future exploitation of the project results following these analyses, the presence of restrictions will be carefully considered. However, I would like to emphasize that third-party data which I plan to use during the project are already publicly available in online databases (e.g. GEO, STRING).

3. Data storage and preservation
3.1 How will your data be stored and backed-up during the research?

The data will be stored in the IT USI-EOC facility, with a capacity of several TBs. The data will be collected in dedicated folders, with limited access only to laboratory personnel and regular backups will be performed daily with an automatic backup program by IT service.

3.2. What is your data preservation plan?

I will retain all the data (raw and elaborated) at least until their full exploitation (e.g. publication of related scientific papers).
Images will be stored as .tif.
Data in spreadsheets will be stored as .xlsx or .csv.
Data in freetext documents will be stored as .txt.
These formats are standard and should support future access and reuse.
Any data which has to be stored in a proprietary format will have the necessary software (including version number) noted in the associated info.txt file.

4. Data sharing and reuse
4.1. How and where will the data be shared?

Transcriptomic, proteomic and secretome data will be available without restrictions and will be shared on adequate public repositories (e.g. Gene Expression Omnibus) for reuse by

other researchers. I will mention this possibility of access in the scientific articles describing the results of the project.

4.2. Are there any necessary limitations to protect sensitive data?

Data will be available as soon as the articles describing them will be accepted for publication.

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Biomedicine #2

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

We will collect data from a prospective cohort of 302 patients with myocardial infarction (MI)-related left ventricular thrombosis (LVT). A screening log for all MI patients with LVT admitted in each center and reasons for not inclusion will be also collected. We will collect the following data from routine clinical assessment: medical history, physical examination, conventional baseline 12-lead ECG, standard transthoracic echocardiography, contrast echocardiography and/or cardiac magnetic resonance imaging (MRI). We will collect imaging and clinical data at baseline and at follow-up. In addition to the randomization visit, there are six scheduled follow-up visits, at 30 days, 3, 6, 12, 18 and 24 months after randomization. Clinical events will be evaluated by a blinded independent clinical events committee (CEC) of 3 experienced physician that will adjudicate adverse events according to standard definitions. Imaging follow-up (with the same qualifying imaging modality, either contrast TTE or cardiac MRI) will be performed at baseline, 6 and24 months and earlier if further clinically indicated by persistence of LVT. A dedicated core-lab, blinded to the treatment assignment, will collect anonymized imaging scans and perform the qualitative/quantitative analysis blinded to treatment arm. Imaging data (echocardiography, cardiac MRI) will be stored in DICOM format in a central institutional storage. All data will be anonymized. Data will not include genetic data. Clinical data that are collected for each research subject constitute the research CRFs (Case Report Forms) including demographics, cover clinical data, laboratory analyses results, medications, and clinical events.

1.2. How will the data be collected, observed or generated?

An electronic data capture (EDC) system will be built for the study. Only encoded data will be captured. The EDC system will include electronic case report forms (eCRFs) designed to capture study information, which are completed by trained site staff. All eCRFs should be completed in a timely manner, preferably within 5-10days of the subject’s enrolment or follow-up visit. eCRFs documenting SAEs should be submitted via the EDC system within 24 hours after the investigator becomes aware of the event. Data images will be sent to a central core-lab that will provide planned analysis. Each local investigator is required to prepare and maintain adequate and accurate case histories designed to record all observations and other data pertinent to the investigation on each individual treated with the investigated medicinal product. All required data will be accurately recorded by authorized personnel on eCRFs. The training of appropriate clinical site personnel to complete the eCRF will be the responsibility of the study monitor. The investigators will keep a patient file containing the source documents (patient informed consent, medical history, laboratory analyses, medication lists, investigational drug delivery records, follow-up visits, AE/SAEs, etc.). The investigator will also keep a patient identification list with complete identification information on each subject (name, address, contact number, pseudonym). This documentation will be kept in strict confidence and will only be accessible to the investigator and study staff. Upon request, these documents must be available for monitoring as well as for inspection purposes. All these documents will be kept for a period in accordance with national regulations.
The Local Investigator is required to prepare and maintain adequate and accurate case histories designed to record all observations and other data pertinent to the investigation on each individual enrolled in the study.
A GCP-compliant EDC system will be used for the study. The subject’s anonymity will be maintained and the confidentiality of records and documents that could identify subjects will be protected, respecting the privacy and confidentiality rules in accordance with applicable legal requirements. Patient data will be encoded:

- Subjects will be identified only by their assigned study number, initials and year of birth on all CRFs
- The investigator will keep a Patient Identification List with complete identification information (name, address, contact number) on each subject.
- The investigator will maintain all study documents in strict confidence.
- CRF entries will be performed by authorized persons and it will be assured that any authorized person can be identified. Copy of ECG, 2D TTE, contrast TTE or cardiac MRI will be stored at the main site.

All data will be cleared of any sensible personal information; patients will be identified by their assigned study number. For end-point adjudication data will be examined without any form of identification (blinded).

1.3. What documentation and metadata will you provide with the data?

The scientific results will be presented at research meetings and published in research journals. The research protocol will be registered in a study protocol repository (e.g. clinicaltrials.gov) before study initiation. Study protocol will be made available for publication with the main manuscript in the supplementary appendix of the article. Study data will be made available to other researchers upon reasonable request and permission. A scientific summary of the significant output of the research, as well as a lay summary, will be generated.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

Protocols used in this study will be approved by local Ethical Committee for Clinical Research. Only eligible subjects will be enrolled. Any research subject is entitled to withdraw from the study at any stage. He/she is also entitled to require a complete removal of his/her data from the study database upon request. Data will remain accessible for authorised people throughout the study conduct, and until the final publication of results. Even though research subjects will be anonymised, consent of research subjects is mandatory in order to enroll them in the study and collect their data. No personal data or data that may easily identify subjects will be provided, with respect to the Swiss law on human research (Federal Act on Research involving Human Beings).

2.2. How will data access and security be managed?

Physical access to the electronic data centers is locked and limited to authorized personnel using personal ID and password authentication. Passwords will need to be changed every 42 days. Remote access to servers is reserved to authorized personnel. VPN access is possible by using two-factor authentication based on etoken and windows domain account. Account lockout policy is active. Only institutional email addresses will be accepted for any communication regarding sensitive data. Per-Tests (simulations of walware attacks) are regularly performed. Personal accounts are granted individually for each person. Identification is made by a personnel ID and a password. Failure to provide the correct password after a limited number of attempts automatically deactivates the faulty account (protection against non-authorized attacks).

2.3. How will you handle copyright and Intellectual Property Rights issues?

Primary owner of all data collected in this study are the subjects that participate herein. In order to use their data for scientific research, approval by the subjects and by an ethical committee will be been obtained.

3. Data storage and preservation
3.1 How will your data be stored and backed-up during the research?

All clinical digital files will be collected and stored on central servers for clinical research data. All data are backed up from the experimental computers to the departmental file server and database server. The file servers and database servers are backed up onto magnetic tapes overnight, and stored in a fire safe. Project data will be stored on file servers managed by IT service in folders with limited and managed access permissions. The principal investigator (applicant) will be in charge of deciding which researchers have access to the folders containing the data during the project (project team) and after, especially if some researchers will change their institutions. Internal Users are using separate windows domain accounts to access files and network resources, using strong password and forced to change it every 42 days. In addition, account lockout policy is active. Backups operations are performed by the IT services, in accordance with the clinic investigation unit policies. Frequent backups are performed using the best enterprise backup solutions and are physically stored in a fire-proof safe. Backup strategy compromises an optimised hourly, daily, monthly and yearly retention plan. Daily backups will be performed every day and data will be copied on tapes.
For security reasons, the application tiers and the database management systems run independently in separate servers. All systems and applications are continuously monitored. Appropriate measures are automatically taken whenever an alert is issued.

3.2. What is your data preservation plan?

The project will archive the entire database in a reusable format. Archives encompass all raw data, metadata, transformed data, transformation operations, deviations, version history, and audit trails will be stored for at least five years after the publications of the project's results. Data will be provided to authorised third parties as much as possible in non-proprietary formats (text, CSV, XML, PDF).

4. Data sharing and reuse
4.1. How and where will the data be shared?

Making data or material available to third parties is ethically and legally restricted by the content of the written and informed patient consent. Individual research subjects' data cannot be made available to non-authorized people. Relevant results related to the research will be presented at national and international conferences and will be published in scientific peer-reviewed Journals.
Study data and metadata will be made available to other researchers upon reasonable request. A scientific workshop will be organized at the end of the fourth year of activity to discuss study results.

4.2. Are there any necessary limitations to protect sensitive data?

Legal restrictions to reuse of the data are based on the Federal Act on Research involving Human Beings (HRA) of Switzerland including the Individual research subjects' data cannot legally nor ethically be made available to non authorised people. Only the sponsor, the investigation team, reviewers, auditors and inspection authorities are entitled to access such data.
No personal data or data that may easily identify subjects will be provided, with respect to the Swiss law on human research (HRA).

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Biomedicine #3

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

The data collection during this study entails clinical, biological, and imaging data as described in the proposal. The clinical data will be stored physically in a secured access to a web-based database, conceived by the means of FileMaker Pro11, will be created. The central server will be managed by the database manager of the institution and will be located at the institution. Principal Investigator will supervise the electronic database central server management. Regarding clinical data, all research subjects will be anonymised. Doubleblinding will be applied for the study team and the patients through the duration of the study. Data that are collected for each research subject include following areas/topics:

- Inclusion criteria and consent
- Demographics
- Medical history
- Concomitant medication
- Laboratory analysis
- Medical events
- Hospitalisation and interventions
- Adverse events
- Imaging outcomes
- Questionnaire

Part of the data will be obtained during medical examinations of the subjects. This includes laboratory data as results of examination of biological samples. Part of the clinical data (retrospective study cohort) will be drawn out of the database already existing at the host institution. Imaging data (echocardiography, CT, MRI) will be stored in DICOM format in a central institutional storage.

Partner institution #1 (Partner1) contribution: The partner will generate cellular patch clamping data (sharp electrode recordings) of atrial tissue from rabbits (e.g., action potential shape and duration). All data will be obtained at baseline and after acute or chronic exposure to sex hormones estradiol. progesterone, or dihydrotestosterone. All data will be stored in digital form, in the format in which it was originally generated (ABF files), and in Microsoft excel, Prism Graphpad, PClamp, and Origin for data analysis.

Partner institution #2 (Partner2) contribution: The partner will produce 1. digital data and 2. collect blood and tissue samples (not strictly regarded as ‘data’ but nevertheless potential source of information for reuse).
- Digital data: Electronic case report forms of recruited patients (eCRF), electrograms-files, genomic data, data of next-generation sequencing (mRNA expression in cardiac tissue samples, raw and processed data), data on histology and targeted biochemical analysis, biomarkers results, algorithms of prediction models. In all data types raw and processed data will be generated.
- Blood samples and tissue samples: Blood serum, blood plasma, buffy coat, all frozen and stored at -80°C. Tissue samples, frozen and stored at -80°C.

1.2. How will the data be collected, observed or generated?

All clinical, and imaging data will be generated and then pseudonymized at the study center where all data are acquired. Each dataset will first undergo quality control before it will be entered into the central database. Quality control will include assessment of completeness of data (no missing information) and adherence to the pre-defined CT / MRI protocol (CT and MRI technical parameters are available in the DICOM header information of each scan).The project members will use an Electronic Data Capture (EDC) System to enter data electronically.

Partner1 contribution: All rabbit experiments included in this project will be conducted including appropriate controls to ensure validity. Data consistency will be assessed by comparing repeated measures. Quality of analytical data will be guaranteed through calibration of devices, intra- and inter-assay quality controls, repetition of experiments and comparison with literature and previous data. Files will be named according to a preagreed convention. The dataset will be accompanied by a README file that describes the directory hierarchy. Each directory will contain an INFO.txt file describing the experimental protocol used in that experiment. It will also record any deviations from the protocol and other useful contextual information. Image files will be stored with a range of metadata that specifies the precise conditions used to obtain each image. This should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future.

1.3. What documentation and metadata will you provide with the data?

Raw data will be provided with its full description, data-types and encoding (comprehensive Code Book).
Metadata regarding database implementation, audit trail, data status, completeness, modifications and intervention dates will be available for each single value. A complete description of the whole project and its technical configuration, as well as a complete version history of structure alteration/modification to the database implementation will be made available upon request. All deviations will similarly be logged in.
Any post-processing of data will be fully described by the description of all transformation operations.
Metadata will be provided as much as possible in non-proprietary formats (text, CSV, XML, PDF).
However, some material may remain proprietary due to technical restrictions. Relevant additional data material not present within the main CDMS database will, whenever necessary, be attached to the raw data, and will be fully described.
Metadata will form a subset of data documentation that will explain the purpose, origin, description, time reference, creator and terms of use of a data collection. Due to the different nature of the research data generated, the metadata will be based on a generalized metadata schema including elements such as:

- Title: free text
- Creator: Last name, first name
- Date:
- Subject: Choice of keywords and classifications
- Description: Text explaining the content of the data set and other contextual information needed for the correct interpretation of the data
- Format: Details of the file format
- Identifier: DOI
- Access rights: open access

It will be the responsibility of:

- each researcher to annotate data with the appropriate metadata,
- the Principal Investigator to check regularly with all participants to assure data is being properly processed, documented, and stored.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

Even though research subjects will be anonymised, consent of research subjects is mandatory in order to enroll them in the study and collect their data. No personal data or data that may easily identify subjects will be provided, with respect to the Swiss law on human research (Federal Act on Research involving Human Beings).

Approval from the competent authority and from an appropriately constituted Competent Ethics Committee (CEC, e.g. CCER) is mandatory before conducting the study. Only eligible subjects will be enrolled. Any research subject is entitled to withdraw from the study at any stage. He/she is also entitled to require a complete removal of his/her data from the study database upon request. Data will remain accessible for authorised people throughout the study conduct, and until the final publication of results.

Partner1 contribution: Atrial tissue from wild type rabbits will be used in these studies. The necessary ethical authorizations for conducting the animal experiments in this project will be obtained prior to the start of the experiments from the proper Cantonal Commission. In performing the experiments, Partner1 strives to strictly adhere to the 3Rs principle of Replacement, Refinement, and Reduction. All researchers and technicians working with the animals receive proper animal welfare training in conformity with DFE Ordinance 455.109.1 on 'Training in animal husbandry and in the handling of animals'.

2.2. How will data access and security be managed?

Medical data is considered as sensitive personal data, regardless to pseudonymization, by the Federal Act on Data Protection. Archiving of the data is regulated by the Federal Act on Data Protection and the Human Research Act of Switzerland. With regard to privacy rights the required period of archiving should not be prolonged. Exceptions need to be explained. Physical access to the data centers is logged and limited to authorised personnel using badge authentication. On a regular basis, vulnerability testing is performed to reduce potential exposure.
Remote access to servers is limited to authorised personnel. Connections to servers are encrypted using SSH. System logs are stored in a dedicated centralised system for audit purposes. Only people part of the investigation team, as well as inspection authorities are given access to data. Personal accounts are granted individually for each person. Identification is made by a personnel ID and a password. Failure to provide the correct password after a limited number of attempts automatically deactivates the faulty account (protection against non-authorized attacks).
Only institutional e-mail addresses will be accepted for any communication of sensitive data regarding account creation and management. Per-Tests (simulation of malware attacks) are regularly performed, and measures taken whenever necessary.

Partner1 contribution: Regarding rabbit data, this will be stored on the institutional centralized file storage system, managed by the IT department of the institute. The central storage facility has redundancy, mirroring and is monitored. For data transfers, exports of all or any kind of partial data will be systematically password encrypted before being transferred. Use of hashing encoding will ensure that no data alteration may have occurred during the transfer.

2.3. How will you handle copyright and Intellectual Property Rights issues?

Primary owner of all data collected in this study are the subjects that participate herein. In order to use their data for scientific research, approval by the subjects or by an ethical committee has been obtained as explained above.
All experiments data will be stored together with metadata including information on the person that recorded the data, the research group and the project. All data will be associated with a unique identifier (DOI) so that the data can be used for other publications without violation of copyright / intellectual property rights.

3. Data storage and preservation
3.1 How will your data be stored and backed-up during the research?

All data and applications are physically stored in dedicated data centres on the principal institution premises. The physical hardware consists of enterprise-grade servers, networking, and storage with trustworthy and stable GNU/GPL solutions.

All applications are hosted on certified virtual servers. There are several dedicated servers for each system, ensuring separation between the testing / pre-production environment and the production environment. System updates are applied only after validation in the pre-production environment. For security reasons, the application tiers and the database management systems run independently in separate servers. All systems and applications are continuously monitored. Appropriate measures are automatically taken whenever an alert is issued.
Backups operations are performed by the IT services, in accordance with the clinic investigation unit policies. Frequent backups are performed using the best enterprise backup solutions and are physically stored in a fire-proof safe. Backup strategy compromises an optimised hourly, daily, monthly and yearly retention plan. Daily backups will be performed every day and data will be copied on tapes.
The tapes will be are moved offsite every month in an institutional warehouse outside the HQ and branch office. No specific identification data, such as patient's surname and name, will appear in the database.

Partner1 contribution: For the runtime of the project rabbit experiments data will be stored on the centralized file storage system of the institute, managed by the IT department of the institute.
The central storage facility has redundancy, mirroring and is monitored. In addition, the data will be regularly backed-up on password-protected external hard-drives.

Partner2 contribution:
- For Digital data: Genomic and next-generation sequencing data processing, analysis, and storage will take place on an HP ProLiant server dedicated to large-scale data analysis. All other data will be stored on an institutional server. The computational infrastructure is housed in the institution's principal IT facility and is maintained by a professional system administrator. The system features a full RAID configuration for internal backup and the research data will also be backed up through the institution's IT system. The server is located behind a firewall; external access is only possible through the use of a VPN tunnel. All clinical digital files will be collected and stored on central servers for clinical research data. The data will be periodically transferred to a server provided.
Data management and storage will be subject to monitoring. Following approval of this proposal, the details of the data management, storage, and sharing strategy will be worked out in collaboration with a Research Data Management Service which provides support for hardware (secured data servers), data storage and access tools as well as support for legal and ethical issues.
- Blood samples and tissue samples: All blood samples and tissue samples will be temporarily stored at the institutional biobank.

3.2. What is your data preservation plan?

At the end of the project, the entire database will be archived in a reusable format. Archives encompass all raw data, meta-data, transformed data, transformation operations, deviations, version history, and audit trails. Redeployment of the entire database is therefore possible whenever needed. The comprehensive archive will be preserved during a minimum period of 10 years. Data will be provided to authorised third parties as much as possible in non-proprietary formats (text, CSV, XML, PDF).

Partner1 contribution: Since the volume of experiments data remains in a tractable range, there is no necessity to delete any data, and all data from completed projects will be preserved / archived on dedicated external drives in departmental archive. There is no legal obligation to destroy data. Data of highest relevance and potential value for re-use (e.g., data from successful experiments on which publications are based) will be transferred to a repository as described under 4.1.

4. Data sharing and reuse

4.1. How and where will the data be shared?

Only metadata that describes the content of the study data base will be made available to the third parties to evaluate if the data of this study fits their intended use of the data. The data will be presented as coherent database, single data sets linked by the pseudonym used during data collection. This data will be sufficient to answer scientific questions deviating from the scope of the Original study protocol, e.g. within meta-analysis.

Partner1 contribution: EUDAT will be used to share experiments data because it supports the FAIR principles (https://eudat.eu).
The DOI issued to datasets in the repository can be included as part of a data citation in publications, allowing the datasets underpinning a publication to be identified and accessed.

Partner2 contribution:
- For Digital data: All digital data (raw or processed) will be stored in the long-term in a redundant RAID configuration on the institutional server. In accordance with funding organization and journal requirements, data will be placed in public repositories while ensuring patient confidentiality protections.
The data will remain accessible to the research team of Partner2 even after the funding period of this SNF project.
- For Blood samples and tissue samples: After the funding period of the study, all blood and tissue samples will further be stored at the institutional Biobank at least until all relevant data have been published. The exact time period for storage still has to be determined but likely will be in the range of a few years.
All data will be stored and made available for reuse by applying the FAIR principles which state that data have to be findable, accessible, intraoperable and reusable.
Two main instruments for implementation of FAIR principles will be used:
1. The use of data repositories: Data repositories offer access to a defined circle of authorized users and have become a broadly used way of data sharing. Data repositories can be local or the respective software is publically accessible. For example, Dataverse (https://dataverse.nl/) is an open-source data repository software installed in dozens of institutions globally to support public community repositories.
Dataverse generates a formal citation for each deposed data set following a defined standard. Dataverse generates a Digital Object Identifier, or other persistent identifiers (Handles) which are made public when the dataset is published (‘Findable’). This digital object identifier is citable and points to a landing page, providing access to metadata, data files, dataset terms, waivers or licenses, and version information, all of which is indexed and searchable (‘Findable’, ‘Accessible’ and ‘Reusable’). Deposits include metadata, data files, and any complementary files (such as documentation or code) needed to understand the data and analysis (‘Reusable’). While this data repository is very broad there are also repositories that are used for specific types of data. For example, the NCBI gene expression omnibus is a common repository for RNA sequencing data (https://www.ncbi.nlm.nih.gov/geo/). Likely all of the options mentioned above will be used for the data sets, depending on the type of data.
2. Open access publications: Following the official publication strategy of our institution all papers and reports will be published open access. All publication will contain references to the respective repositories providing access to the data.
Importantly, making data or material available to third parties is ethically and legally restricted by the content of the written and informed patient consent. The applicant will design the patient information and written consent forms such that later reuse of data will be possible. Nevertheless, before reuse can take place the applicant will investigate together with the requesting party whether the reuse is covered by the patient consent and whether it complies with the relevant regulations.

4.2. Are there any necessary limitations to protect sensitive data?

Legal restrictions to reuse of the data are based on the Federal Act on Research involving Human Beings (HRA) of Switzerland including the Individual research subjects’ data cannot

legally nor ethically be made available to non-authorised people. Only the sponsor, the investigation team, reviewers, auditors and inspection authorities are entitled to access such data.
No personal data or data that may easily identify subjects will be provided, with respect to the Swiss law on human research (HRA) and its applicable ordinance ClinO/KlinV/OClin/OSRUm.

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Communication #1

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

1. Data that will be reused:
- In agreement with CERN, text and visual sources inside their archive (period 1989-1993) will be digitized (PDF format). We will need at least 10GB of storage space.
- Sources retrieved in newspaper and private archives. The materials will be digitized, when available only in paper format. These archives are in part public and in part privately owned. Access will be required for private or semi-public archives. We plan to digitize around 2GB of data.
- Primary-sourced histories from the founders available in specialized books and websites. It will be both text files and audio files. The text files will be digitized, when available only in paper format, or downloaded, if already available in digital format. The audio files will be downloaded when possible. Audio and digitized texts will be stored in the project folder on USI servers. All these materials are publicly accessible.
Audio: 50MB per each file (around 350MB in total).
2. Data that will be generated during the project:
- Digitally recorded interviews with the Web pioneers about the birth and early years of the Web: duration of each interview around 1 hour; 25-30 interviews to be saved in .Wav format. Around 200MB per file (6GB in total)
- Transcriptions of interviews: 25-30 PDF.

1.2. How will the data be collected, observed or generated?

Materials collected in the different archives will be digitized through scans of text and images (output format: PDF. Files) The files will be named as follow: ArchiveName_FolderName _Year_Month_Day _FileNumber.
The files will be saved in folders pointing to the reference year within the original archive. For each archive, information regarding the files will be collected in an Excel file containing general information (Year, File name, Type of document, details of content, comments).
The interviews will be downloaded and transcripted. Both files will be named as follows: NameOfInterviewee_Inverviewer_Date_FileNumber.
An excel list of the interviews containing general information (Name of interviewee, Date, Description Comments) will be provided.
Generated data will be transferred as soon as possible to the USI servers to avoid data loss.

1.3. What documentation and metadata will you provide with the data?

For each archive, information regarding the files will be collected in an Excel file containing general information (Year, File name, Type of document, details of content, comments).
An excel list of the interviews containing general information (Name of interviewee, Date, Description Comments) will be provided.
Content analysis will produce a catalogue of PDF files pertaining each document collected within the archives and containing technical data (metadata about the document) and the results of the thematic analysis.
Data from the content analysis will be later on processed through SPSS and an excel sheet with the main results will be provided as well.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

We have a special permit to access, process and preserve classified materials at the CERN archive. Collected data won't be made public without CERN agreement. Data collected remain property of CERN.

Permission to access semi-public newspaper archives and private archives will be asked in due course. Data collected within these archives won't be made public without the agreement of the archives owners and managers. Data remain property of their respective owners.
Informed consent will be signed by the interviewee. Interviews will be made available for research purpose. Investigated themes are not sensitive, therefore we do not provide anonymization of data. Anonymization will be provided if explicitly required by the interviewee.
Interviews included in books or websites are already publicly available data and will be treated as such.

2.2. How will data access and security be managed?

Data collected within the different archives and digitized will be accessible only to the research team and will be stored on USI servers.
There are no specific security needs concerning the collected data. Project data will be stored on file servers managed by USI IT service in folders with limited and managed access permissions. The project manager will be in charge of deciding which researchers have access to the folders containing the data.

2.3. How will you handle copyright and Intellectual Property Rights issues?

Data collected within the different archives remain property of their respective owners. They will not be shared without the owners' permission. Data generated by the project (interviews and content analysis results) are property of the Università della Svizzera italiana.

3. Data storage and preservation
3.1. How will your data be stored and backed-up during the research?

The project data will be stored and processed on file servers managed by the IT service of USI, which are protected by backup copies of data (backup) made every night. Cloud services are not intended to be used. Researchers will be trained to transfer data from the recording devices to the file servers as soon as possible to avoid risks of data loss.

3.2. What is your data preservation plan?

Data collected within the CERN archives and digitized by the research team will be preserved until the CERN archive will be opened to the public. The excel list of the materials available within the archives will be preserved for future research.
Data collected within the semi-public or private archives and digitized by the research team will be preserved until these archives will be opened to the public. The excel list of the materials available within these archives will be preserved for future research.
PDF files produced by the content analysis will be preserved until the final results of the research will be published. The excel list with the summary results of the content analysis will be preserved for future research.
Interviews digitally recorded in .wav format will be preserved together with PDF transcriptions.
10 years after the end of the project data will be re-evaluated in order to determine what to preserve.

4. Data sharing and reuse
4.1. How and where will the data be shared?

Subject to agreement with the researched archives, the list of the documents collected will be made available on Zenodo for future research purpose. Excel summary of the results produced by the content analysis will be made available on Zenodo for future research purpose. Detailed PDF files and SPSS documents produced by the content analysis will be made available upon request and only for research purposes. A written agreement will be

signed in order to ensure the proper usage of the data made available to other interested researchers.
Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citable. We will assign relevant titles and keywords to the datasets, so that researchers who are interested can easily find them.

4.2. Are there any necessary limitations to protect sensitive data?

Data collected within the different archives remain property of their respective owners. They will not be shared without the owners' permission.
Data generated by the project (interviews and content analysis results) are property of the Università della Svizzera italiana and will be made available for research purposes only.

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Communication #2

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

In order to allow for a comparative analysis as foreseen in this project, we will have four different, interrelated and (probably) multilingual datasets: (a) 40-50 interviews to women in the Swiss academic system and their partners, corpus I in the research plan, (b) 40-50 individual interviews to women in the Swiss academic system, corpus II in the research plan; (c) 15-20 interviews to key institutional actors in the field of equal opportunities and (d) written public institutional documents about equal opportunities; (c) and (d) together make up corpus III (institutional discourse) in the research plan.
Datasets a, b, c (interviews) will be collected during the project. Based on previous similar experience, we anticipate that each interview will last about 45-60 minutes, which amounts to a total recording time (sum of datasets a, b, c) between 71.25 and 120 hours of recording. This requires between around 43GB and 75GB of storage for MP3 audio files. The data collection of datasets a, b will be preceded by a short pilot study that will enable us to refine the interview questionnaire and also to make more precise calculations about the length of interviews and the required data storage.
In a first phase, interviews will be first transcribed in a non-anonymized form, only available to the researchers as Word files (and deleted at the end of the project). The interviews will then be anonymized. Anonymized interviews will be stored in two versions. First, Word/PDF versions will be created for the purposes of the project (publications, research) and for future sharing of data (see also 4.1). Second, XML versions will be created and annotated using the freely available software UAM Corpus Tool (CT, http://corpustool.com, see O'Donnell 2008), version 3.3. CT has already proven useful for the purposes of an argumentative analysis and has been already used by applicant and post-doc in previous projects. CT requires transcriptions to be in XML files.
At the moment, around 80 public institutional documents in English, Italian, French and German (as PDF files) have been collected and stored on USI server, in a folder dedicated to this project. Presently, the public institutional documents require about 58MB of storage; the collection of documents, however, will be regularly updated during the project and we estimate an increase of at least 15-20 documents over the four years.

1.2. How will the data be collected, observed or generated?

Audio recordings will be made using digital recorders that are made available for researchers at USI by IT services, which allow to do recordings of the quality required for this project. Interviews will be recorded, respectively, by the two PhD students (datasets a, b) and by the post-doc researcher (dataset c). The post-doc researcher also has the task to work on the collection of public institutional documents (dataset d). During the pilot study, and during the process of data collection, the applicant and the post-doc will work with the PhD students to refine the methodology and give feedback on possible difficulties emerging during the data collection; during regular weekly project meetings, any urgent issue can be addressed by the project team.
Files of the datasets a, b, c, d will be stored separately on three USI IT folders (see 2.2). New versions of interviews/documents will be clearly labelled; the post-doc will have the task of managing the process of updating versions.
Transcriptions of interviews will be helped by two student assistants. This will allow double review of data, which enable us to enhance the quality of transcriptions.

1.3. What documentation and metadata will you provide with the data?

For datasets of interviews (a, b, c) we will create metadata to allow the project team and other users to understand and reuse the files. We will create an identifying name (dataset letter + number of interview) for each file and corresponding transcription. For datasets a, b,

in a separate Excel file, we will insert metadata: time/place of interview, language(s), name of interviewer, plus basic sociological data on interviewees, such as age, education, number of children, nationality, time spent in the Swiss academic system, level of academic career, year of completion of the PhD. In case of a couple's interview, sociological data will also include: family status of interviewee, relationship to the interviewed person (e.g. partner/friend/relative).
For dataset c, in a separate Excel file, we will insert metadata: time/place of interview, language(s), name of interviewer, plus basic sociological data on interviewees, such as age, education, number of children, role within equal opportunities in Switzerland (e.g. head of service/administrator/other). At the moment, we have labelled and sorted the documents with a first classification but the metadata will be refined during the project.
We use Excel files for our metadata because, although the software UAM Corpus tool permits to insert meta-data, it is likely that we will start the annotation after all interviews have been transcribed; we want to compile the Excel files while progressively storing and transcribing the interviews.
Also concerning datasets a, b, c (interviews) in a separate Word file called JOURNAL, the PhD students and post-doc will write up a short comment (in the form of a journal) immediately after each interview, noting possible ambiguities and everything that was not clear to them during the interview. From previous experience, we know that this additional information will be useful when analysing data. This file will remain available only to the project team because it will include confidential information.
PhD students will be trained about metadata from the applicant and post-doc and they will be continuously monitored during the project. Two student assistants will help with the transcriptions but providing precise metadata is the PhDs' responsibility, under the general monitoring of post-doc and applicant. Note that the metadata described above are for strict use of the project team. A public version of the Excel file of metadata, in which only non-sensitive information will be included, will be created to make transcriptions available to other researchers (see section 2.1 below).
For dataset d (public institutional documents), we will create metadata to allow other users to reuse the files. We will create an identifying name (dataset letter + number of file) for each file. In a separate Excel file, we will insert metadata: name of institution that issued the document, website from which the document was downloaded, time when the document was downloaded from the website language, available translations.

2. Ethics, legal and security issues
2.1 How will ethical issues be addressed and handled?

Institutional documents (dataset d) have been downloaded from public websites of Swiss higher education institutions.
For individual interviews (datasets a, b, c), including the pilot studies, all participants will sign an informed consent, compliant with GDPR. They will be informed that only transcriptions of the data in an anonymous form will be published and possibly shared with other researchers, only for scientific reasons.
The applicant will get support from the USI ethical committee to define the legal and ethical issues for this informed consent, as she did with previous projects. Given the focus of this project, all participants will be adults and therefore will have the right to sign their own informed consent. The participants will be informed that they can withdraw their permission to use the data. They will be given the applicant's and researcher's email addresses, and know that they can always contact the project team, if they want to have further information.
As said in 1.1., we will produce as soon as possible an anonymized version of the transcription, removing names and all identifying information (Word/PDF/XML). For this anonymized version, only a selection of metadata (language, age of the interviewee, number of children) will be made public. Excerpts of interviews can be published only for scientific reasons. Audio files and other metadata will only be available to the project team in order to assure the anonymity of the participants. The "JOURNAL" Word file will remain available only to the project team because it will include confidential information.

Confidentiality is important because the interviews, especially to women academics, may include sensible topics and, in order for us to access the data we need, we need participants to feel they can express themselves freely.
Other possible legal and ethical issues emerging during the project will be discussed by the project team under the supervision of the applicant and, if needed, with USI ethical committee.

2.2. How will data access and security be managed?

Project data will be stored on file servers managed by USI IT service in folders with limited and managed access permissions. The principal investigator (applicant) will be in charge of deciding which researchers have access to the folders containing the data (see also 2.3) during the project (project team) and after, especially if some researchers will change their institutions.

2.3. How will you handle copyright and Intellectual Property Rights issues?

USI - Università della Svizzera italiana will be the owner of the data. Audio files and sensible metadata will only be available to the project team in order to assure the anonymity of the participants. Only anonymized versions of the transcriptions (PDF) will be made public to other researchers (limited by a confidentiality agreement) 5 years after the end of the project.

3. Data storage and preservation
3.1. How will your data be stored and backed-up during the research?

The project data will be stored and processed on file servers managed by the IT service of USI, which are protected by backup copies of data (backup) made every night. We do not intend to use cloud. Researchers will be trained to transfer data from the recording devices to the file servers as soon as possible to avoid risks of data loss.

3.2. What is your data preservation plan?

At the end of the project, the non-anonymized transcriptions of interviews will be immediately deleted. All other data will be kept in the USI folders, for further research, maintaining the standard of confidentiality used throughout the project. We will keep the audio files because they could be useful for further publications of the project team. After 7 years, the applicant will have the task of reconsidering what to do with the audio files – whether to keep them for another 5 years or longer, or to delete them, depending on data value.
The annotation made in UAM Corpus Tool will also be a profitable basis for future research on argumentation; therefore, we will keep the XML files. PDF files of anonymized interviews (not annotated) will be shared more broadly (see 4.1) five years after the end of the project, while annotated anonymized versions will be shared seven years after the end of the project (see 4.1).

4. Data sharing and reuse
4.1. How and where will the data be shared?

As soon as they are ready, anonymized metadata will be made publicly available on the repository Zenodo (https://zenodo.org), which is a non-commercial repository. Our datasets will likely be multilingual so we will specify that on Zenodo.
The audio data stored at USI will only remain available to the applicant and members of this project; while copies of annotated interviews (XML files annotated with UAM Corpus Tool) could be made available during the project, if needed for research purposes, to other researchers who sign a confidentiality agreement – it is the applicant's responsibility to make this decision. Five years after the end of the project, anonymized transcriptions of the interviews in PDF format will be made available on Zenodo only for research purposes in the field of argumentation and discourse analysis, to researchers who sign a confidentiality agreement. Seven years after the end of the project, annotated interviews (also anonymized)

will also be made available on Zenodo only for research purposes in the field of argumentation and discourse analysis, to researchers who sign a confidentiality agreement. Dataset (d) is composed of written public documents taken from websites of universities. While its metadata will be shared, the documents per se will not be shared on Zenodo. Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citable. We will assign relevant titles and keywords to the datasets, so that researchers who are interested can easily find them.

4.2. Are there any necessary limitations to protect sensitive data?

Initially, only anonymized metadata will be made publicly available on Zenodo. Anonymized transcriptions of the interviews in PDF format (not annotated) will be made available only for research purposes in the field of argumentation and discourse analysis five years after the end of the project. Seven years after the end of the projects, the anonymized annotations will be made available only for research purposes on Zenodo. In order to further protect sensitive data, we will anyway ask to sign an agreement to those researchers who are interested to use our anonymized datasets making sure that they will only use the data for research purposes.
We will be very strict when interpreting what it means to produce "anonymized" transcriptions, checking that any information that could make the interviewee identifiable will not be shared. The applicant will monitor this aspect throughout the process, asking relevant advice from the Ethical committee at USI when necessary.

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Economics

1 Data collection and documentation
1.1 What data will you collect, observe, generate or reuse?
The project will combine the following datasets:
- Equity returns and other stock-market information of publicly listed companies
- Balance-sheet information of these companies
- Macroeconomic indicators

Part of the data is obtained from data providers (Datastream, Orbis) for which USI has an existing licence. Other data (such as, some macro series) are from central banks and other organizations (such as the FED or World Bank). Data is received in CSV or Excel format. We anticipate that the overall dataset will amount to approximately 20 GB.

1.2. How will the data be collected, observed or generated?

The equity data will be accessed and downloaded from Datastream and Obis. Other data will be downloaded from the website of economic organizations (such as the FED or World Bank). To keep track of possible revisions/versions of the data, they be stored into a folder that is named based on the date of download. To guarantee the replicability of our analysis, new downloads are stored in separate folders, and previous downloads are preserved.

1.3. What documentation and metadata will you provide with the data?

The data will be accompanied with a precise description of the series that are used in the study, along with their reference code when present (e.g., Mnemonic in Datastream). The raw data will then be processed (through some dedicated software such as Stata, Matlab, or Phyton) to be converted into a format suitable for the empirical analysis. The procedure used to process the data will be described in a technical report or in a data section of the publication, including the type of software used, and made public together with the file and an indication of the software version that was used.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

There are no ethical issues in the generation of results from this project.

2.2. How will data access and security be managed?

Data will be stored on a dedicated disk space of one of the USI servers that are managed by our IT department. The data will be accessible using virtual desktop technology only by authorized participants to the project. The list of authorized participants will be managed by the applicant. Access to the database will be logged, thus each access is traceable.

2.3. How will you handle copyright and Intellectual Property Rights issues?

Part of the data is obtained from data providers (Datastream, Orbis), and is subject to a non-disclosure agreement. Other data (such as, some macro series) are from central banks and other organizations (such as the World Bank) and are freely available. The research is not expected to lead to patents.

3. Data storage and preservation
3.1. How will your data be stored and backed-up during the research?

Data will be stored on a dedicated disk space of one of the USI servers that are managed by our IT department. Our servers have redundancy, mirroring and are monitored. The servers are backed up on a regular basis (at least once per week).

3.2. What is your data preservation plan?

We will preserve the data for at least 10 years on the university's server, and also deposit it in an appropriate data archive (such as Zenodo). Where possible, we will store files in open archival formats, such as Excel files converted to CSV. In case this is not possible, due to the large volume of the data, we will include information on the software used and its version number.

4. Data sharing and reuse

4.1. How and where will the data be shared?

Datasets from this work which underpin a publication will be made available through the project participants' websites or common published on Zenodo, and made public at the time of publication. Data in the repository will be stored in accordance with funder's data policies. The retention schedule for data will be set to 10 years from date of deposition in the first instance, with possible extension for datasets which remain in regular use.

4.2. Are there any necessary limitations to protect sensitive data?

Datasets from this work which underpin a publication will be made available at the time of publication. Data that is purchased by third-party data providers will be published in a form that does not constitute a violation of the license agreement.

4.3. All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4. I will choose digital repositories maintained by a non-profit organisation.

Yes

Informatics

1. Data collection and documentation
1.1. What data will you collect, observe, generate or reuse?

Data will be collected from static and dynamic program analyses applied to well-known benchmarks and to code in public software repositories. Depending on the concrete metrics to be collected (which will be investigated in WP 1 and WP 2) and the number of open-source projects where our analyses are applicable, the total amount of data collected may reach the order of terabytes. We will maintain proper scripts to be able to reproduce the results of our evaluations.

1.2. How will the data be collected, observed or generated?

We will use scripts to fully automate the collection and aggregation of measurements, to compute statistics, and to generate figures conveying the results of our experimental evaluations. We will pay special attention to the reproducibility of our results. To ensure that experiments are repeatable on specific workloads, we will save detailed provenance data, such that one can analyse a given workload version even if it is superseded by a more recent version. The maintenance of provenance data is particularly important for large-scale evaluations on open-source projects in software repositories such as GitHub. However, please note that full reproducibility of experiments is only possibly for deterministic workloads.
For metrics that are prone to fluctuations, we will repeat measurements a sufficient number of times and report mean values as well as different quantiles, variance measures, and confidence intervals. In some cases, measurements need to be taken after a warmup phase of a system (e.g., when a system has reached a steady state after initial dynamic compilation); the details of the measurement procedures will be maintained in scripts to ensure consistency of measurements and reproducibility of results. Measurement results will be kept in private repositories shared by the project team. We will use free and open-source distributed version control systems such as Git.

1.3. What documentation and metadata will you provide with the data?

For each data set, we will store metadata about the input data (e.g., concrete revisions of open-source software) and on the toolchain (e.g., version numbers of used analysis tools) used to obtain the data. The details of the measurement process will be provided via scripts that allow other researchers or practitioners to repeat the measurements and reproduce the data. The scripts will be properly documented to ease their reuse.

2. Ethics, legal and security issues
2.1. How will ethical issues be addressed and handled?

Our planned research does not raise any ethical concerns.
For the evaluation of our methods and techniques, we will be working on open-source projects maintained in public code repositories. The code and data contained there are freely available and so our results do not need any particular protection, as we are not dealing with any sensitive data.

2.2. How will data access and security be managed?

As explained before, we will not be dealing with sensitive data. Hence, the results of our work do not require any special protection.

2.3. How will you handle copyright and Intellectual Property Rights issues?

The software developed in this project will be released open-source under the most appropriate license to benefit the community. Thus, the researchers funded by this project can freely exploit the results of their research in their future careers.

3. Data storage and preservation

3.1. How will your data be stored and backed-up during the research?

The project team will use backed-up, private repositories (such as those provided by the university) for sharing documents, code, data, and any other artefacts during the project. Completed and tested software will be released open-source in public repositories. Data collected in experiments will be kept in a distributed database. Since we are not dealing with sensitive data and the results from experimental evaluations can be easily recomputed, special precautions regarding data storage are not needed.

3.2. What is your data preservation plan?

As discussed before, all relevant artefacts will be kept in backed-up repositories, ensuring their long-term availability. Completed software will be released open-source and maintained in public code repositories. We also plan to submit artefacts for accepted publications, which will be made available through the publisher's digital library.

4. Data sharing and reuse

4.1. How and where will the data be shared?

During the research activities, all data and artefacts will be shared between the project team (and possibly with external collaboration partners) in backed-up, private repositories (e.g., provided by the university's IT department). Completed and tested software will be released open-source and kept in public code repositories to ensure the widest possible impact and reuse. Experimental data underlying publications will be submitted as artefacts, to be included in the publisher's digital library. We will also include scripts to reproduce all measurements and the figures included in our papers.

4.2. Are there any necessary limitations to protect sensitive data?

Data underlying a publication will be publicly released as an artefact accompanying the publication. Whenever supported by the publisher, we will submit artefacts with our papers for inclusion in the publisher's digital library. Scripts used to reproduce measurements will be part of the open-source releases of our software, to be maintained in public code repositories.

4.3 All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4 I will choose digital repositories maintained by a non-profit organisation.

Yes

Contacts

Research and Transfer Service
Università della Svizzera italiana
Via Buffi 13
6900 Lugano
Switzerland

e-mail igor.sarman@usi.ch
web https://www.usi.ch/it/universita/info/srit