Research Data Management Policy SBE

Version: 1.2
Date: 20 February 2017

General Introduction

Good data stewardship is one of the prime responsibilities of a professional research organization. Vrije Universiteit has adopted a university wide[1] data management policy, accompanied by faculty specific supplemental policies. The remainder of this document details the data management policies of SBE.

Data management is a multi-facetted concept. The focal points of the current policies are primarily:

- ensuring scientific integrity and accountability, and in particular research verifiability;
- ensuring research compliance with the legal environment, in particular on privacy sensitive data;
- ensuring data access over longer periods (10 years), even in cases where researchers leave the university.

SBE also fosters the concepts of open access, open data, and reproducibility (which requires a much higher level of standardization and user-friendly documentation of research output). These additional concepts, however, are not part of this policy document.

Further updates of this policy document are expected as data management requirements and standards, legal restrictions, and IT infrastructure develop further.

Aims, scope and responsibility

1. The policies regarding data management and data stewardship at the School of Business and Economics (SBE) as laid out in this document explicate (to the faculty level) the general data management policies of Vrije Universiteit Amsterdam.[2]

2. The aim of these policies is to help SBE researchers ensure that:
   - they meet legal and ethical data management requirements, including privacy of study subjects;
   - they meet the requirements of research funders;
   - they meet the requirements of scientific journals concerning quality and traceability of data;
   - they protect SBE's and their own scientific integrity.

> *Data stewardship refers to the long-term and sustainable care for research data. Data stewardship implies professional and careful treatment of data throughout all stages of your research project (i.e., the design, collection, processing, analysis, long-term preservation, and sharing of your research data).*
>
> Data4Lifesciences, 2016

[1] See VUNET, search for "data management", and then the link to the general VU policy.
[2] See VUNET, search for "data management", and then the link to the general VU policy.

3. The policies in this document apply to all research by SBE affiliated researchers **initiated after January 1, 2017 and aimed at resulting in a publication in a peer-reviewed journal, a book, or a book chapter.**[3] For the remainder of this document, the expression "published research" or "research" is meant in the sense defined above.
   A proviso holds for research that builds on previous research for which the data gathering process was initiated before January 1, 2017, and for which not all steps in the current SBE data management policies can be retraced. Here researchers should comply with the policies on documentation and archiving as far as is reasonably possible. Researchers should always comply with legal constraints (e.g. for privacy sensitive data).
   Notwithstanding the above, all SBE researchers are invited to comply with all data management procedure for their research initiated before January 1, 2017, that has not yet been published. In all cases, researchers should comply with legal constraints (e.g. for privacy sensitive data).

4. Each individual SBE researcher is responsible for adhering to the policies in this document and the general VU data management policies to conduct sound data stewardship. In case of doubt, researchers should consult a senior colleague for expert advice. If doubt persists, a specific question should be raised towards SBE's scientific committee via SBE's research policy officer (i.putter@vu.nl).

Storing and documenting the data and the data gathering process

5. All research data shall be **secure, compliant, and adequately documented**. This policy encourages **reusability** and even **shareability** of research data, but does not enforce this.

6. The quality of the stored data complies with the required standards in the *Nederlandse Gedragscode Wetenschapsbeoefening*. This includes the requirements that
   (i) all steps in the research process can be checked and should (in principle) be replicable, and
   (ii) that the quality of gathering data, data input, data storage, and data processing are monitored and controlled well.

7. Data and the data gathering process should be well documented, such that both are (in principle) verifiable. The documentation should preferably be part of the published research itself, or of (online supplemental) appendices of the published work. If not, the researcher should provide detailed supplemental README files and archive these securely along with the data. The documentation on the data gathering process should be sufficiently detailed and at least include:

> *Adequate data stewardship ensures that*
>
> - *you have adequate technological resources (e.g., storage space, support staff time);*
> - *you are able to share your final data set publicly;*
> - *your data will be robust and free from versioning errors and gaps in documentation;*
> - *your data is backed up and safe from sudden loss or corruption;*
> - *your data will remain accessible and comprehensible in the future;*
> - *your data can be shared with others, for scientific research, commercial development, or validation.*
>
> Data4Lifesciences, 2016

[3] Effectively, this includes all outlets that could qualify for research time in SBE's internal annual research output assessment, and more. It covers the entire complex diversity of research cultures and data needs at SBE, such as experimental research, surveys, commercial databases, proprietary internal data, externally managed databases, qualitative and quantitative data, big real-time data sets, etc. Future versions of these policies may extend the scope of application further, for example to published working papers.

- details on the process of gathering the raw data (including survey data, interviews, video material, experiment scripts, details/code on how websites were scraped, etc., if appropriate);
- detailed **metadata**, including descriptions of the variables (possibly also with database tickers/acronyms in case outside databases are used, and including details on the interpretation (e.g., does 1 indicate male or female as a gender dummy));
- details on filters and data manipulations used to get from the raw data to the data used for the empirical analysis (including the removal of outliers or individuals from the original sample (e.g., on Trial 5 of Participant 10 there was other student interference and hence the responses are replaced by NaNs), operations on variables (winsorizing, trimming, scaling, rotating), details on how recorded interviews were transcripted and coded, etc.);
- ethical clearance (if needed; see SBE's Research Ethics policies);[4]
- details on how privacy issues are dealt with (if personal data are involved);
- details on how access to the data is arranged and ensured (at least for internal purposes) for the minimum period of 10 years.

8. Notwithstanding that most funding organizations require[5] ex-ante a written **data management plan**[6] for grant approval, the current SBE policies encourage, but do not require an *ex-ante* written data management plan for all its research.

9. For ongoing research, data should be stored securely and professionally. It should not be possible that the loss of a single data carrier or the departure of an individual researcher makes it impossible for the rest of the research team (at VU or elsewhere) to retrieve the data. Researchers should make sure that back-ups of the data and their accessibility are properly arranged and communicated within the research team.
   Responsibility for ensuring this rests with the principal researcher of the project. Details on the arrangements are known to at least two people in the department at any time, one of whom may be the departmental secretary.

10. Upon publication, researchers archive their raw[7] and processed research data, unless compliance issues arise (e.g., license issues), or storability becomes an issue (for some big data sets; see later in this document). Archiving means: to store data on a secure system, *together with* the data documentation (see point 7 above). To comply with VU policies, research data that are not relevant for further research need to be archived securely for a period of 10 years.[8] Data that

[4] See http://www.feweb.vu.nl/nl/onderzoek/research-ethics-integrity/feweb-research-ethics-review-board/index.aspx.

[5] See for instance http://www.nwo.nl/beleid/open+science/datamanagement.

[6] For a template of a data management plan, see VUNET, data management plan (also in Annex 1).

[7] What precisely is labeled as the raw data may be a trade-off between efficiency of storage and verifiability, particularly in qualitative research processes. For example, transcripts of interviews or their coded versions may sometimes be classified as the raw data, rather than the original audio recordings. Similarly, the scraped internet information may be classified as the raw data, rather than snapshots of each of the underlying websites that was scraped.

[8] Data that are relevant for re-use are preferably archived for a longer period. Research data that may be relevant for future research and that may need to be stored longer includes: (1) unique data: Data that cannot be collected again. For example: the operating temperature of today; (2) data of scientific or historical value: research which reflect a period in history, such as an interview with a veteran of World War II or photographs from a certain period; (3) valuable data for re-use: Data that could be valuable to other researchers so that they do not need to collect or structure the data themselves. Such data would ideally be stored indefinitely.

cannot be archived, e.g., due to legal constraints, should still (in principle) be verifiable by archiving the data documentation (point 7 above).

11. Ideally, data (including details on the data gathering and manipulation process) are archived on the journal's website together with the original publication, such that lasting access is ensured of both the publication, the data, and the data documentation.[9] If this is not possible, secure archiving possibilities should be found that ensure data accessibility for the required minimum period of 10 years, also in cases where the individual researcher is no longer available. Researchers, in coordination with their head of department, should up-front the relevant funding for this, if needed.

12. Secure storage and archiving is meant to imply that there are measures preventing data loss (using back-up facilities and proper hardware maintenance) and data leakage. Possible (fee-based) acceptable institutional solutions are:
    - **DataverseNL**, an online platform for the analysis and publication of research data in a semi-open environment. DataverseNL allows users to link directly from publication to dataset and to share it.
    - **ArchStor**, a research data archive with a 10-year retention period. Data stored in ArchStor can only be accessed for verification purposes.
    - **DarkStor**, an offline archive for storing sensitive information/data like privacy or copyright. Once archived, access to the data can only be requested by authorised individuals, i.e. the original researcher or a research coordinator.

    More information can be found on VUNET.[10]
    Other storage options may be possible as long as the data remain available for 10 years. This excludes storage options where the individual researchers pays a fee for the storage facility to continue its services: in this case the storage depends on the availability of the individual researcher for the period of 10 years. Such facilities may only be used if the financial commitment is transferred to the researcher's department.

13. Data materials and carriers are stored as much as possible in a digital format (not in paper format). File formats and other standards should be suitable for long-term preservation and accessibility, e.g.: .pdf, .txt, .csv, etcetera.[11] Also for picture or movie material the most long-term resistant format should be used. All data files contain information (or accompanying README files) about the software that is needed to open the file (and which version of the software has been used).

14. Data owned by the researcher should be licensed to Vrije Universiteit for the minimum duration of 10 years for at least purposes of research verification.

[9] For example, Elsevier has introduced the "Data in brief" option upon acceptance of a paper. Similar initiatives are expected elsewhere.

[10] For example, go to VUNET and search for one of the above database names.

[11] Note that data files of standard software such as SAS, STATA, SPSS, Matlab, etcetera, but also well used formats such as xls and .xlsx or .docx are less robust as they are not always transferrable from one version of the software to the next. The simplest formats should be used.

Privacy sensitive (personal) data

15. Personal data is privacy sensitive and warrants special attention, as there are substantial legal requirements for this type of data. Personal data are data relating to (e.g. passport ID number) or (potentially) traceable to (a combination of birth date and zip code) to a person.[12] Particularly sensitive are race, religion, and health.
    Researchers who collect and process personal data in their research have a legal requirement to notify the Data Protection Officer (DPO) of VU. This is a legal obligation imposed by the Dutch Personal Data Protection Act. Notification can be sent by a standard form in VUNET: search for "persoonsgegevens", or go to
    https://vunet.login.vu.nl/services/pages/detail.aspx?cid=tcm%3a164-414602-16
    If in doubt whether your data is legally classified as personal data, contact the VU's Data Protection Officer (DPO) at servicedesk.privacy@vu.nl.

16. To assess the risk of personal data, the researcher will ask the Data Protection Officer whether a Privacy Impact Assessment (PIA) is needed to assess the institutional risk of the data, and the required security level for storage. ITVO (IT for research) can help with the implementation of this within the VU cost allocation model.

17. For sensitive data, also additional security measures are needed for storage. The researcher is aware of the document "Guideline for working with personal information for scientific purposes" in VUNET on the Data Management page. The researcher and his research group are responsible for data that can put society or SBE at risk.

18. Personal data must at all times be stored on a secured network and in accordance with legal guidelines. PCs should always be locked if the researcher is absent.
    Sensitive data should never be stored on unprotected data carriers (including unencrypted hard-drives in password protected[13] computers) or on synchronized cloud services (Dropbox, Google drive), particularly if these mirror to local non-encrypted hard-drives.
    Again, we refer to the document "Guideline for working with personal information for scientific purposes" for further details.
    After the research, the data can be archived on a safe, off-line server such as DarkStor to comply both with safety requirements and the minimum archiving period.

19. Researchers should, if possible, anonymize the data before sharing the data or storing the data. Keys relating the anonymized data to the personal data should be kept safely and separately and be accessible by at least three persons affiliated to Vrije Universiteit at all times, including after the departure of the original researcher. It is advised that next to the researcher, at least the principal researcher and two departmental secretaries (of related departments) have access to the keys, and that these keys are stored on at least two restricted locations, either physically or electronically. The locations for key storage are either physically (safe) or electronically (H: drive

[12] We refer to https://autoriteitpersoonsgegevens.nl/nl/over-privacy/persoonsgegevens (in Dutch) for precise legal details and definitions.

[13] The hard drive could be removed and copies to be read on another machine.

or other secure medium) sufficiently well secured.
The database ID number of individuals cannot contain any potentially meaningful identifier (such as initials, date of birth, postal code).

20. Encryption is a preferred further security measure.[14] Additional safety measures should meet the general requirement stipulated above, that data access should not be lost in case of the loss of a data carrier, the departure of an individual researcher, or the loss of one of the keys. In particular this means that encryption keys are available to at least two persons affiliated to Vrije Universiteit at all times, including after the departure of the original researcher, and that data and keys are stored in at least two locations, both with strongly restricted and regulated access.[15]

21. After the research has been completed and / or the data archiving period has elapsed, all data carriers with personal data should be deleted in accordance with legal requirements.[16]

22. If a researcher wants to share personal data owned by VU with outside parties (including co-authors), the researcher should make sure to comply with the legal requirements, inclusive of the compliance of his co-author. The VU legal office has drafted a “bewerkingsovereenkomst” for this purpose. The researcher should consult with the VU Data Protection Officer (DPO), servicedesk.privacy@vu.nl, to be fully briefed on the requirements to share personal data outside VU.

23. Personal details should only be archived for more than 10 years if there are reasons to continue the research and only if there is explicit consent from the participants for archiving these data.

24. If a VU researcher processes personal data owned by other parties (e.g., ministries, corporates, banks, etc.), the researcher is responsible for safe storage as described above. The SBE researcher should make sure that his/her role as “bewerker” is well documented in the “bewerkingsovereenkomst” that is typically issued by the owner of the data. Again, the VU Data Protection Officer (DPO), servicedesk.privacy@vu.nl, should be contacted in case of any doubt.

25. For use of personal data of non-EU countries, additional / other rules may apply. The researcher is responsible for compliance with such rules and to keep contact with the VU Data Protection Officer (DPO), servicedesk.privacy@vu.nl.

26. Researchers should not **travel** with privacy sensitive data in unencrypted format. In a number of countries importing encrypted data is illegal and should be avoided (e.g., Russia, China[17]).

[14] VU offers True Crypt for these purposes.

[15] It is advised that next to the researcher, at least the principal researcher and two departmental secretaries have access to the keys, and that these are stored on at least two restricted locations, either physically or electronically.

[16] Additional guidelines will be added here to help researchers. Coordination with the legal office of VU is ongoing.

[17] For some of the challenges, see for instance http://security.uri.edu/travel/travel-to-china-or-russia/.

Scripts, codes, and particular data types

27. Researchers archive their **codes and scripts** used to produce their research. This includes C-codes, Matlab, SPSS, SAS, STATA, Eviews scripts, etc., including a README file on the version of the language or package in which the codes were run.[18]

28. SBE researchers engaging in **experimental research** should take extra steps in gathering and documenting their data. In particular, they should document their correspondence with SBE's Research Ethics Review Board[19] if such correspondence is required and also make sure that they obtain and archive the active consent of their respondents. Respondents have the opportunity and means to withdraw their active consent any moment in the future and leave the experiment from that moment onwards. Withdrawals should be noted in the data documentation.
    For further guidelines, the faculty policies follow the policies of the top journals on experimental research, such as:
    - Economics: https://www.aeaweb.org/journals/policies/data-availability-policy or https://www.eeassoc.org/index.php?site=JEEA&page=42&trsz=40
    - Marketing: http://www.ejcr.org/PDFs/Research_Integrity_Policies.pdf.

29. **Secondary data** are data that are collected by for instance a third party for their own research, data that are collected by companies internally, or data collected by institutions that specialize in data collection (such as CBS, Eurostat, FED, Datastream, Reuters, Bloomberg, etc.). If the data provider allows for archiving the secondary data, this is preferred. However, secondary data need not be archived if the data are (in principle) recoverable (possibly after paying a fee or establishing the appropriate contacts). The data documentation (including the process of gathering and constructing the raw and cleaned data) should still be archived and contain sufficient information on how the data were obtained and accessed, and details (and possibly scripts) to get from the raw (possibly proprietary or commercial) data to the data used for the empirical analysis, similar as in the standard case. This includes the mentioning of company contacts if the data were obtained through private contacts or if agreements with the company disallow the local archiving of the data at Vrije Universiteit.

30. **Big data.** Some data sets are too large for (standard) storage. Here researchers will use the best practices in their field, and share these explicitly with the faculty's research officer (i.putter@vu.nl) such that the faculty can develop a more concrete policy in this area. In all cases, including Big Data studies, the data gathering and construction process should be documented and archived and contain sufficient details such that the published research can (in principle) be verified.

Hardship clause

31. Any exceptions of the above data policies shall be decided upon by the SBE board upon advice of the SBE's scientific committee.

[18] Though not enforced by the current policy document, researchers are encouraged to provide clean and readable versions of their code.

[19] Research Ethics Review Board: http://www.feweb.vu.nl/nl/onderzoek/research-ethics-integrity/index.aspx and http://www.feweb.vu.nl/nl/onderzoek/research-ethics-integrity/feweb-research-ethics-review-board/index.aspx.

Annex 1: Data Management Plan – Template (Concept 1; VUNET 11 Nov 2016)

Section 1: General outline	
Version:	
Date:	
Research Project Title:	
Head researchers:	
University Faculty & Department(s):	
Co-operating organization(s):	
Grant organization(s):	
Grant number(s):	
Contact details	
Address:	
Telephone:	
Email / website / Facebook:	
Summary / Description:	
Secondary data sources:	
Section 2: Responsibilities	
Summary description:	
Operational gathering & storage:	
Data analysis & documentation:	
Lead programmer(s):	
Meta-datastandard(s) used:	
Quality Control manager:	
Section 3: Technical description:	
Summary description:	
Software:	
Hardware:	
Network/cloud facilities:	
Section 4: Legal framework	
Legal owner(s) of primary data used:	
Legal owner(s) of secondary data sources used:	
Applicable law or legislation:	
Description of security / access restrictions to datasets:	
Confidentiality agreement :	Yes / No
Summary description of relevant confidentiality agreement:	
Are there contractual limitations?:	Yes / No
Summary description or relevant contracts:	
Ethical / Privacy issues?:	Yes / No
Summary description of ethical / privacy issues:	
Relevant scientific protocols used:	
Section 5: Publication / Archiving	
Summary description of publication requirements:	
Summary description of long-term storage method:	
Date of storage:	
Legal storage-term:	. . Years
Names of organization(s) & departments involved:	

Contact details	
Scientific/technical (storage) standards used:	
Meta-data standards used:	
Create Commons (CC) License	Yes / No
Data Access restrictions:	Open access / On Request / Restricted
CC License description:	
Embargo	Yes / No
Summary description of embargo:	
Section 6: Financial framework	
Summary description of costs involved:	
Software & Hardware costs:	
Personnel costs:	
Long-term data storage costs:	
Total costs reserved:	