Data Management in Psychological Science:
Specification of the DFG Guidelines

*Felix Schönbrodt, Mario Gollwitzer & Andrea Abele-Brehm**

on behalf of the DGPs Executive Board (adopted Sep 17, 2016; current version: Nov 9, 2016)

The present recommendations belong to a number of initiatives launched by the German Psychological Society ("*Deutsche Gesellschaft für Psychologie*", DGPs) to assure and enhance the quality of psychological research. They are borne from the idea of an open and transparent science, in which published findings are reproducible, and data collected in the context of published scientific works and third-party funded projects are made accessible to other researchers for secondary use.

Calls for public access to research data have been ongoing for some time[1]. For instance, in their "Recommendations for Secure Storage and Availability of Digital Primary Research Data" (2009) the German Science Foundation ("*Deutsche Forschungsgemeinschaft*", DFG) demanded that publicly funded data are freely available after the completion of a project[2]. In line with this, in 2010 the Alliance of Science Organizations in Germany called for long-term storage of and generally free access to research data. [3]

In September 2015, the DFG published data management guidelines that affirmed these goals and asked research associations to consider their data management regulations and to develop appropriate standards for discipline-specific use and sharing of research data[4]. The German Psychological Society (DGPs) joins the DFG and the Alliance of Science Organizations in Germany in their mission to specify the DFG guidelines for the field of psychology.

This requires a thorough deliberation of rights, costs, and benefits from the perspective of (1) study participants, (2) researchers who collected the original data, (3) the general public (including potential holders of relevant copyrights), and (4) the scientific community (including potential secondary users of collected research data). The interest of the *scientific community* in the comprehensive use of data must be weighed against the interest of individual researchers to harness the data they collected as well as the interest of study participants in the ethically responsible handling of their data.

The importance of collecting original data in psychology cannot be overstated. Data are a *conditio sine qua non* for any empirical science. Anyone who generates data and shares them publicly should be adequately recognized. Therefore, researchers who collect original data should not face any disadvantage in their career compared to secondary users of data (e.g., because the latter produce a larger number of publications than the former in the same period).

Accordingly, the DFG emphasizes that "the engagement and efforts of scientists to facilitate the availability of research data should be acknowledged when considering their scientific qualifications."[5,6] At the same time, meaningful secondary use of data can lead to important and valid

---

* English translation by Malte Elson, Johannes Breuer, and Zoe Magraw-Mickelson. They original German version of these recommendations is at http://www.dgps.de/fileadmin/documents/Empfehlungen/Datenmanagement_deu.pdf. Many aspects of these recommendations have been suggested by DGPs members during a constructive discussion process. It would take too long to mention every single contributor, but we explicitly thank all contributors for their ideas and their dedication.

1 For a review of the literature until 2012 see Fahrenberg, J. (2012). Open Access – nur Texte oder auch Primärdaten? [Open Access – Simply text or also primary data?] *Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD). Nr. 200/2012*. http://www.jochen-fahrenberg.de/fileadmin/openacces/Open_Access_Primaerdaten.pdf; see also American Psychological Association (2015). *Data Sharing: Principles and Considerations for Policy Development*. URL: https://www.apa.org/science/leadership/bsa/data-sharing-report.pdf

2 http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf

3 http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze.html

4 http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf

5 http://www.dfg.de/en/research_funding/proposal_review_decision/applicants/submitting_proposal/research_data/

scientific discoveries. The aim of the present specification of the DFG guidelines is to implement a balance between those different interests:

- It emphasizes the importance of sustainable research data management,
- it defines what "primary data" are and how they should be stored,
- it defines standards and potential data sharing restrictions and
- it defines the rights and duties of researchers that share data and researchers that use secondary data.

DGPs suggests that external funding agencies may consider these recommendations when deciding on grant proposals and when they review the final reports of research projects[7].

The present recommendations will be evaluated after five years and revised, if necessary. The recommendations are available in German and English (for the German version, see http://www.dgps.de/fileadmin/documents/Empfehlungen/Datenmanagement_deu.pdf).

Editors of journals that are published on behalf of the DGPs may consider the present recommendations when they handle manuscripts to their respective journal.

In addition, the DGPs will make these recommendations available to the international scientific community and work towards achieving a consensus with other international initiatives.

When appointing vacant positions or evaluating the scientific accomplishments of applicants, DGPs members are advised to consider the present recommendations and the extent to which applicants commit to principles of transparency in their own research.

1. Research data management

The aims of sustainable data management in psychology are, among other things, as follows:

a) Quality assurance (guaranteeing long-term verifiability of scientific results, including reanalysis of data with novel or alternative methods that are superior to those available at the time of data collection)[8];

b) Optimizing knowledge (the use of data for reanalysis and meta-analyses, analyses of "unique" data sets)[9];

c) Maximizing the cost-benefit ratio (the optimal use of collected data, avoiding redundant burden for human and animal subjects).

The open, long-term, and free access to research data contributes to achieving these goals. Researchers collecting original data (referred to as "data sharers") must take care in order for the effective use of their shared data (see section 7.2)

Researchers who want to reuse original data for secondary analyses ("secondary users") are obliged to comply with certain standards (see section 7.3).

When sharing data issues regarding protection of data privacy, copyright, and research ethics have to be considered (see section 5). These concerns can impose restrictions on the sharing of primary data.

---

6 One way to further acknowledge the efforts of researchers collecting original data is to create a new publication category, "shared data." In publications based on secondarily used data that are not co-authored by individuals who collected them, the name of the authors of the primary data' and the exact citation of the repository in which these data can be found should be reported close to the publication title. Researchers who share data can include a new category in their personal bibliography (e.g. "Secondary data use by *** in publication ***"), in which publications based on secondary data use are itemized.

7 See also: Hartig, K. & Soßna, V. (2016). *Forschungsdatenmanagement in DFG-Anträgen: Was kann, was soll, was muss beschrieben werden?* [Research data management in DFG grant proposals: What can, should, and must be described?] Institutionelles Repositorium der Leibniz Universität Hannover. DOI: http://dx.doi.org/10.15488/262

8 Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods* [Online]. doi:10.3758/s13428-015-0664-2

9 http://www.allianzinitiative.de/fileadmin/user_upload/redakteur/Grundsaetze_Forschungsdaten_2010.pdf

2. Primary data

The term “primary data” is used repeatedly in the following text; therefore, it is necessary to begin with a definition. First, a distinction should be made between raw data and primary data. *Raw data* are the original record; for instance, checkmarks on a paper questionnaire, drawings, or audio and video recordings. *Primary data* are the first transfer of raw data into a digital format; for instance, code “1” for a “yes”, etc.

Thus, primary data in psychology are completely unaltered (i.e., not transformed, aggregated, etc.) quantitative and qualitative data, for example:

- Each manipulated and measured variable of every experimental session of every study participant in an experiment;
- Each response of every person to every item in a survey;
- Original wording of inputs in free text fields (under consideration of privacy laws, see below);
- Digitized video recordings (note: as these are usually not sufficiently anonymized, they cannot be stored in a public repository. Instead, coding of the observed behavior can be stored);
- Downloads or screenshots of social media content (e.g. Facebook profiles or Twitter messages);
- Transformed (neuro)physiological data (such as EEG or fMRT data) in a standardized raw data format (e.g. EDF, DICOM, or NIFTI) that are not aggregated and not restricted to selected “regions of interest”[10]

“Primary data” also include data of cases that were removed from the analyses (except for cases in which participants withdrew their consent during or after data collection).

3. Storage

Primary data should be made available digitally[11] in a trustworthy repository. The following are important quality characteristics of trustworthy repositories[12]:

- Economic and ideological **autonomy** and scientific **professionalism** of the institutional provider;
- **Persistence** of data: Long-term data storage (at least 10 years, ideally substantially longer) must be guaranteed; there should be a protocol describing what happens to the data in case the repository ceases to exist;
- **Accessibility** of data: It must be possible to retrieve data openly and freely; however, defining access restrictions (in terms of “Scientific Use Files”) should also be possible (for a discussion of optional access restrictions, see section 5);
- **Identifiability** of data: There must be a persistent data identifier (e.g. a persistent URL or, if possible, DOI);
- **Clarification of data property rights**: Storing data must not imply ceding *exclusive* rights of use to third parties (however, *simple* rights of use, i.e. the right to archive and copyright, must be conferred to the operator of the repository);
- The option to store data **publicly as well as non-publicly.**

For these reasons, trustworthy open repositories (e.g. PsychData, ZPID[13], datorium at GESIS[14], or a developed university repository) are preferred over journal repositories. However, storing data on private or personal university websites is not recommended.

---

10 For MRI data see, e.g. the “Best Practices in Data Analysis and Sharing in Neuroimaging using MRI” by the Organization for Human Brain Mapping: http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf

11 It is assumed that the majority of psychological research data is already available in digital format or can be easily digitized. It is recommended to store non-digitizable data at the institution. Researchers are advised to ensure that data storage is maintained in case of an affiliation change or retirement from academia.

12 See also Data Seal of Approval: http://datasealofapproval.org/en/

When choosing a repository, constraints imposed by research ethics guidelines (e.g. prohibition of storage on servers in foreign or non-European countries) need to be considered. The institution that provides the repository service should advise and support researchers intending to store their primary data.

4. Costs of data archiving

Preparing data and making them available in accordance with high quality standards is inevitably tied to an increased resource expenditure. Hence, additional financial support in form of personnel and material resources for the preparation and archiving of datasets can and should be requested in applications for third-party funding. Very large amounts of data (e.g., EEG or fMRI) can be stored in specialized repositories. The use of such repositories entails additional costs that should be included in grant proposals.

5. Data privacy and copyrights

Limitations imposed by protection of data privacy and copyrights need to be taken into account when planning a study.[15] For example, proper anonymization or pseudonymization ensures that individuals cannot be identified by combining various measures, including those collected across multiple studies with the same participants (e.g. first semester psychology students at University XY). [16] Data privacy concerns are not only relevant on the individual level, but also on aggregate levels: Particularly for sensitive research topics (e.g. illegal behavior, suicide rates, etc.), researchers must pay close attention to the extent that institutions (schools, companies, etc.) can be identified by the data or by merging multiple datasets.

Relevant laws and regulations (regarding data privacy and the right to informational self-determination/"Recht auf informationelle Selbstbestimmmung") need to be accounted for at the stage of participant recruitment. This holds true particularly when studying children who cannot give informed consent on their own. Study participants need to be aware that their anonymized data might be made available to third parties for secondary use and that the purpose, nature, and scope of that secondary use is currently not foreseeable. Explicit and specific informed consent for secondary use must be obtained in case data cannot be fully anonymized[17]. When data are fully anonymized, as with survey responses or data from experimental procedures, such specific consent does not necessarily have to be obtained as individuals can no longer be identified. When in doubt, the local ethics committee or the DGPs central ethics committee should be consulted.[18]

Consent forms and ethics application forms should be adapted to comply with these recommendations. Moreover, institutional ethics boards are requested to review whether their own guidelines might be too restrictive with regard to research transparency practices without actually being conducive to the protection of data privacy (mandatory deletion of fully anonymized data, for example, is unnecessary). Suggestions for suitable formulations can be found in appendices B (consent form) and C (guidelines of the ethics committee).

It is imperative to observe legal requirements against the sharing of data where applicable. Not fully anonymized data of individual participants who have denied their consent to potential secondary use must not be shared. If data cannot be shared, an appropriate explanation should be provided (e.g., in

---

13 Dehnhard, I., Weichselgartner, E. & Krampen, G. (2013). Researcher's willingness to submit data for data sharing: A case study on a data archive for psychology. *Data Science Journal, 12*, 172-180.

14 https://datorium.gesis.org

15 See, e.g.: Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. *BMJ*, *340*, c181. doi:10.1136/bmj.c181; or: http://theodi.github.io/ukan-course/#0.1

16 Gola, P. & Schomerus, R. (2010). *Kommentar Bundesdatenschutzgesetz* [Annotation German Data Protection Act], 10th Ed. München: C.H. Beck.

17 Metschke, R., & Wellbrock, R. (2002). *Datenschutz in Wissenschaft und Forschung*. [Protection of data privacy in science and research] Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit. https://datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077

18 Deutsche Gesellschaft für Psychologie (Ed.) (2016). Ethisches Handeln in der Psychologischen Forschung. [Ethical conduct in psychological science] Göttingen: Hogrefe, in Druck.

a footnote in the publication).[19] Conversely, such concerns should not be used as a justification not to share data when it is legally and ethically unproblematic. Further, when legal restrictions to data sharing apply, it should be stated which types of aggregated data or anonymized partial data can be shared.

When data privacy is a concern, "scientific use files" (SUF) restrict access to a specified group of users. SUF require a de facto anonymization (i.e., de-anonymizing the data would be extremely difficult). Such datasets are only shared upon request with researchers at research institutions for scientific purposes. Typically, these datasets cannot be freely distributed. Secondary users have to sign non-disclosure and non-propagation agreements, and are required to delete the raw data after a stipulated period.

The risk of identification should be diligently assessed when dealing with data deserving increased protection, such as self-disclosed ethnic origin, political, religious, or philosophical convictions as well as data on one's health and sex life, and the use of SUF should be considered accordingly. Application of SUF licenses can also be appropriate in cases when an abuse of the data might be anticipated.[20]

The decision tree (figure 1) points out the most significant issues when preparing and sharing anonymized and personal data. Its purpose is to illustrate typical situations. It does not cover all possible cases of how personal data should be handled; particularly with regards to secondary use of personal data, many issues must be considered. We refer the interested reader to the comprehensive guidelines by Metschke and Wellbrock (2002) (see footnote 17).

19 Taichman, D. B., Backus, J., Baethge, C., Bauchner, H., de Leeuw, P. W., Drazen, J. M., Fletcher, J., et al. (2016). Sharing clinical trial data: A proposal from the International Committee of Medical Journal Editors. *New England Journal of Medicine, 374*, 384–386. doi:10.1056/NEJMe1515172

20 Lewandowsky, S., & Bishop, D. (2016). Research integrity: Don't let transparency damage science. *Nature*, *529*(7587), 459–461. http://doi.org/10.1038/529459a

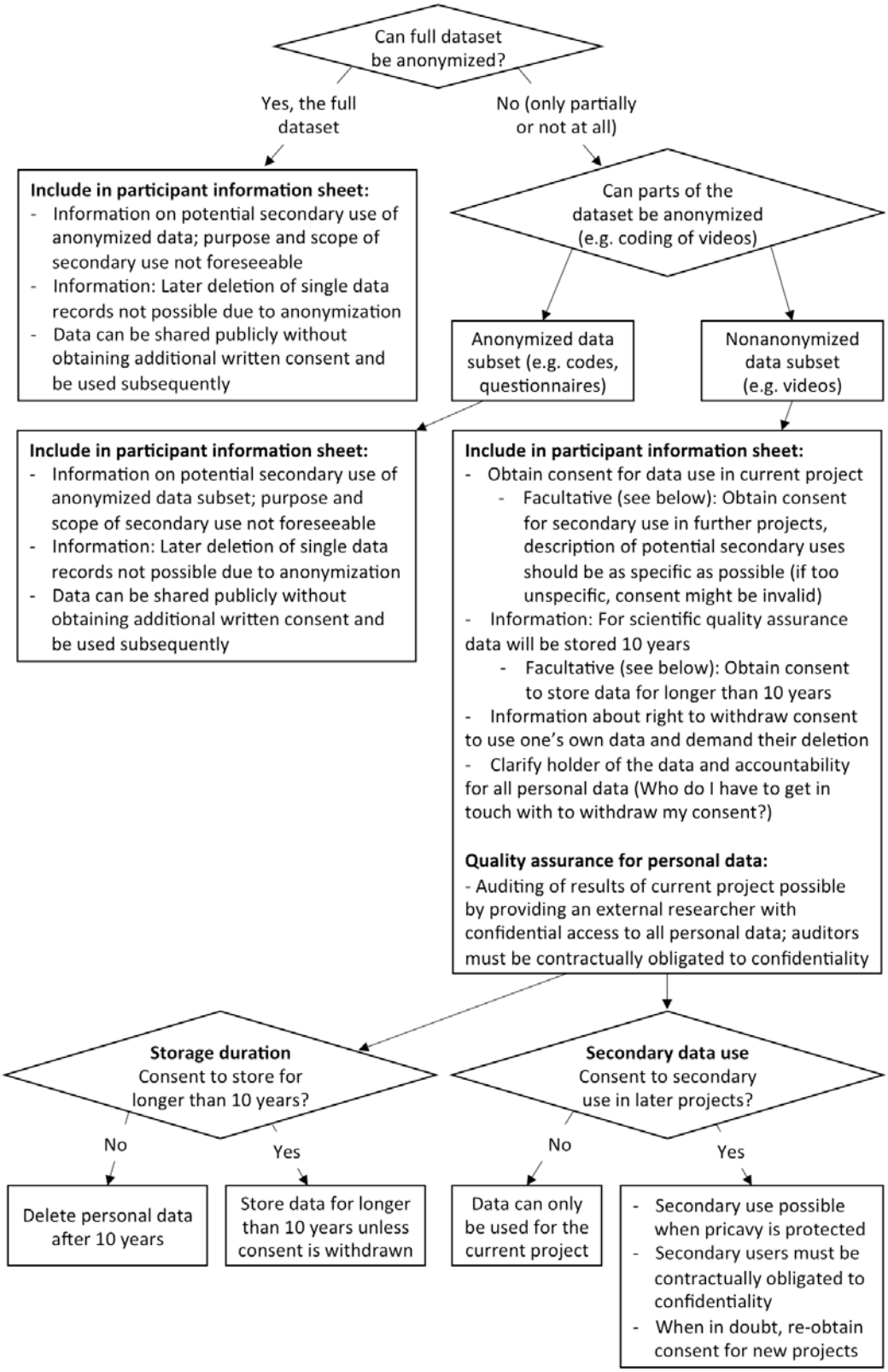

Figure 1: Simplified illustration of the most important issues when preparing and sharing anonymized and personal data.

Not fully anonymized data that cannot be freely accessed online, such as image and audio files, video recordings of persons, interview transcripts (including clinical interviews) etc., should be non-publicly archived for at least 10 years.

The process of anonymizing or pseudonymizing the data should be documented comprehensively. DGPs guidelines on research ethics will soon be published. Other relevant guidelines are already available.[21]

With regards to research transparency, it is recommended to make all instruments available in their original language, unless sharing these instruments would undermine their usefulness or violate other (e.g., copyright) agreements. In the case of commercially available tests procedures, training manuals, and tests for applied settings, professional norms and legal restrictions need to be respected.[22] This also applies to software or similar products for which a patent application or commercial use is intended. Similar sharing restrictions may be imposed for publications based on existing data that require a use license (e.g. SOEP or NEPS[23]). Furthermore, when cooperating with non-German universities, foreign laws should also be taken into account.

6. Time and scope of data sharing

In the following, we distinguish between two types of data sharing.

Data Sharing Type 1: Sharing of data that are part of a publication

With the publication of a manuscript, the person or group who collected the data (the “data sharers”) should make available all primary data necessary to reproduce the published results. This type of data sharing does not only apply to data collected in DFG projects, but more generally to all data on which published articles or reports are based. Unless specified otherwise, the first author or the corresponding author is responsible for providing the documentation and ensuring the reproducibility of the dataset.

A publication should report all other variables measured within the study or studies that were not used for the publication itself (see "standard reviewer disclosure request", https://osf.io/hadz3/). Data from these other measures are only shared when they are used for a publication or when the third-party funded project is concluded and the complete dataset is made available (see below, *Data Sharing Type 2*). In the case of comprehensive survey studies with large amounts of variables (e.g., questionnaire items), reporting a brief overview of the assessed constructs or subject areas accompanied by a link to a more detailed document is sufficient.

Data Sharing Type 1 also applies to research funded by regular institutional resources (e.g., publications resulting from graduate theses or assignments), corporations, as well as non-public third-party funds. Researchers are advised to determine which type or part of their primary data can and cannot be shared for the purpose of reproduction and secondary use before the data are collected. Usually, publications are based on data that can be shared under the Data Sharing Type 1 policy. Exceptions to this (e.g., special agreements with the funding source regarding data sharing) must be declared in the publication.

Data Sharing Type 2: Sharing after project completion

In accordance with the DFG guidelines, the data that have been collected in a funded research project should be “made available to the public immediately after completion of the research or within a few months” (*trans.*).[24] This also includes all relevant data of the project that are not yet part of a publication. It further encompasses all materials (particularly analysis scripts, code books, and – if possible – stimuli) required to make sense of the data. For simulation studies, the data generating code as well as the simulated data should be shared (unless the amount of data exceeds the currently available storage space of repositories). If the materials necessary to replicate a study result cannot

21 http://www.psychdata.de/downloads/PsychData-Handbuch.pdf

22 Sharing commercially available test materials always requires the permission of the copyright holder or publisher.

23 SOEP = Sozio-ökonomisches Panel (German Socio-Economic Panel), http://www.diw.de/en/soep; NEPS = Nationales Bildungspanel (National Educational Panel Study), https://www.neps-data.de

24 see the DFG’s "Recommendations for Secure Storage and Availability of Digital Primary Research Data" (January 2009; http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf)

be made available, a rationale should be provided, for instance, in a README file stored in the repository.

It is at the discretion of the person responsible for the project to decide which data are relevant. Examples of irrelevant data might be those obtained in experiments based on flawed code or in highly exploratory pilot studies. To counter the problem of publication bias, all data from properly conducted studies that did not yield the expected outcome ("null results") must be made publicly available; results not conforming to hypotheses may under no circumstances be suppressed. When final reports are evaluated, reviewers should verify that results and primary data of all proposed studies are available.

The time at which a project is considered completed may vary depending on its complexity and other factors; generally, a project is considered completed when the final report is being submitted. Sharing of the data should occur as soon as possible after the project has been completed.

Under an embargo (see 7.1 and 7.2), data can be stored non-publicly in a repository when the project is completed. This means that the data are uploaded after project completion and already receive a persistent identifier, but are not yet available to the public. After the embargo period has expired, the respective data can be shared freely.

Data Sharing Type 2 applies particularly to projects funded by public (and, if possible, also non-public) organizations and for which both scope and completion are properly defined. For continuously running projects (e.g., with university funds), defining the "completion" might be difficult; however Data Sharing Type 1 applies either way.

7. Rights and duties of primary and secondary users

The idea of an open science, the obligation to share research data, and the possibility of secondary use of shared research data presents challenges both for the sharers as well as the secondary users of data. Both parties have specific rights, but these come with specific duties.

7.1. Rights of data sharers

Researchers who produce primary data (i.e., those who share these data) have the right of first use to the data. If more than one researcher is involved in the project, the rights of first use should be negotiated within that group prior to sharing the data.

Data sharers can define an embargo for secondary use. This means that the data that have not yet been used for publications are stored in a repository, but are not accessible for third parties for a certain period of time (e.g., through password protection or a non-public directory in the repository). This way the data cannot be used by third parties for their own analyses.

Data sharers further have the right to know who uses their data and for which purposes. They have the right to be informed by the secondary users about a secondary analysis of their data prior to publication – especially if the secondary analysis does not reproduce the original results (also see section 7.3 "Rights of secondary users"). Some repositories offer data sharers the option of automated notifications about when their dataset was downloaded and by whom. This *information* about the download of data should generally not entail a *restriction of access* for specific groups of people (except in well-founded exceptional cases; see point 5 under "Scientific use files"). This means that the data sharer is informed about a download through the choice of a suitable repository. Irrespective of this information function, secondary users are *obliged* to inform the primary researchers about any secondary use (whether this is in a publication, a presentation, or a blog post).

7.2. Duties of data sharers

Data sharers are required to share their data in a way that enables a meaningful secondary use. This includes (a) that informed consent about the use of the data from the study participants is available and (b) all data and corresponding metadata that describe the dataset as a whole have to be documented thoroughly and comprehensibly. Tools to facilitate the appropriate data management are currently being developed (e.g., DataWiz by the ZPID).

Data sharers have the right to define an embargo (see section 7.1); in this case, however, data sharers are required to (1) announce this embargo as soon as the initially inaccessible primary data are stored in a repository and to (2) explicitly state the end of the embargo period. This can be done by

way of a publicly accessible file in the repository that describes the embargo, states its end, and describes the collected data (e.g., by referring to a codebook).

After the end of the embargo period, the data will be made publicly available and are normally open to secondary analyses without restrictions, even if the data sharers have not yet used the data for publications themselves.

Experience tells us that with the end of a research project not all planned publications have been finished. As a general rule, the DGPs views an embargo of max. 5 years after completion of a project as adequate. Longer embargo periods need to be justified (e.g., in the file in the repository that also states the end of the embargo).

Generally, researchers should not impose an embargo on data that have been used as part of a publication (Data Sharing Type 1). In exceptional cases, such as when data collection is extremely laborious or if certain follow-up questions for the dataset has already been generated, then researchers can also impose an embargo on these datasets. However, this embargo should be considerably shorter than the embargo defined for Data Sharing Type 2. In addition, it has to be ensured that these data are also available upon request for reproduction of the reported analyses once they are published.

7.3. Duties of secondary users

In order to guarantee the advantages of secondary data use while, at the same time, minimizing its risks, transparency, trust, and the willingness to cooperate are essential for all parties involved in the process. Secondary users should contact the data sharers to facilitate a valid use of the data and to avoid misunderstandings. Any use of data should be guided by the principle of maximum knowledge advancement in relation to a research question. Accordingly, the secondary use of data should not be motivated by the aim of damaging the reputation of the data sharers. Conversely, data sharers must not prevent the publication of the results of a reanalysis that contradict the original findings or reveal errors in the original work.

Especially if secondary users intend to make their reanalysis public (e.g., in presentations, publications or blog posts), the data sharers have to be informed (1) that the data are used and with what aim, (2) the results of this secondary use, and (3) where the results of the secondary use will be published.

In any case researchers who use data shared by others have to cite the data adequately.[25] For this purpose, it is helpful if the data are accompanied by a citation reference (including a persistent identifier) in the repository. A publication that is based on secondary data use should not share its own version of the dataset, but always refer to the persistent identifier of the original dataset, even if the dataset has been changed in the course of the reanalysis (e.g., by computing new variables). Potential data transformations or newly computed variables have to be documented in reproducible analysis scripts.

Moreover, secondary users are required to analyze the data in a way that does not infringe the rights of the participants of the original study. The secondary use of data is subject to the same data protection and copyright laws as the primary use. It is the duty of the secondary users to ensure that these are adhered to.

Of course, the secondary use of data has to meet the same requirements with regard to transparency and scientific diligence as the primary use. The scientific standards for a reinterpretation of the data have to be those that are valid at the time of the secondary analysis. For the evaluation of the original analyses in the context of the reanalysis, the scientific standards that were valid at the time of the original analysis have to be applied.

Offering co-authorship. The question whether or under what circumstances data sharers should be offered co-authorship on a publication resulting from a reanalysis of the data cannot be regulated in general and has to be answered case by case. The question who made a substantial contribution within a project that warrants co-authorship for a publication also has to be posed in the case of secondary use. However, we propose the following categorization that shows some (although in no way exhaustive) examples of co-authorship in secondary data use:

25 Data Citation Synthesis Group (2014). *Joint declaration of data citation principles*. https://www.force11.org/datacitation

- *Simple secondary data use*: for instance, extracting effect sizes for meta-analyses; computing means or distributions of variables. For this type of secondary use, the data sharers are typically not offered co-authorship. Reanalysis that exclusively attempt to reproduce the original results (e.g., to be reported in a blog post) generally belong to this category of simple secondary use.
- *Extended secondary use*: we distinguish between two subcategories here:
    - *Additional questions that complement or expand the research question of the original publication*: for instance, when the reanalysis of an available dataset shows that the main published effect is moderated by another variable that was measured; in a follow-up study this moderation is tested in a confirmatory fashion. Thus, the theory of the original authors is further conceptually refined. In such cases, the data sharers should be offered co-authorship.
    - *Orthogonal analyses which use the data to answer a different question than the original publication*: such as when a researcher develops a new measure of reliability and applies it to a variety of available survey data from different researchers (that were focused on non-methodological questions). For this type of secondary use, it is not necessary to offer the data sharers co-authorship.

In case of doubt in a particular case, the original authors should always be contacted. Co-authorship should also always be offered if the contribution of the data sharers to the secondary use goes beyond the mere data sharing. An example of such a contribution could be additional advice and support for the reanalysis of the data.

If a secondary use goes beyond the scenario of simple secondary use described above, we recommend that the data sharers and the secondary users come to an agreement that also regulates the question of co-authorship. Such an agreement could also address questions of data protection. Additionally, an agreement could regulate in what way the data sharers may comment on the results of the secondary use (i.e., if they are not co-authors on the resulting publication). An agreement of this kind should, however, not be used to selectively exclude certain (groups of) researchers from secondary use. If the data sharers have defined an embargo, this has to be respected. Any violation of a valid embargo (or of any other contractual agreement) means that the secondary data analysis must not be published or that respective publications must be retracted.

Appendix E illustrates the different data sharing types and possibilities of secondary use using an example.

The rules for secondary use presented here can also be enclosed in the repository (e.g., in a Readme file), so that secondary users can familiarize themselves with them, particularly if were unfamiliar with the present recommendations.

Appendix A: Illustrative Examples

The DGPs is aware of the fact that the open data sharing brings along new requirements that – until now – have often not been heeded in the research process. Hence, the DGPs wants to provide recommendations so that data sharing can meet the high quality standards of research transparency. In addition, the DGPs offers workshops that provide a practical introduction to this topic.

1. Storage duration: The DFG specifies that primary research data have to be stored for *at least* 10 years. This regulation is derived from good scientific conduct that specifically aims at avoiding scientific misconduct and ensuring that research findings can be reproduced and verified in case of doubt. This does not imply that the data *ought* to be deleted after this period. As there is no argument for the deletion of anonymized data – with the exception of studies with extremely large amounts of data – the DGPs typically advocates an unlimited storage without time-dependent deletion mechanism or obligatory deletion.
2. Storing the data in a nonproprietary format (e.g., .csv or .txt files) should be preferred over proprietary formats (e.g., .mat, SPSS or SAS files) to avoid limiting transparency to users/owners of potentially expensive specialized software. Accordingly, data should be prepared in a way that allows them to be read without the need for special software. A binary output from an unknown experimental software would not meet the goal of research transparency. In such a case, data have to be prepared so that they are in a comprehensible data format. If this is not possible, there should be a documentation that explains with which software can read the data.
3. In addition to the technical accessibility the readability with regards to content has to be ensured for the data. All variables must be documented in a digital codebook. [26] It must be clear which manipulated or measured variable in the publication belongs to which variable in the dataset. Ideally, the dataset is supplemented by analysis scripts (e.g., R scripts or SPSS syntax) that reproduce the published results.
4. The storage location should receive a persistent identifier (e.g., a persistent URL or – if possible – a DOI). This allows for a consistent citation of the data. The storage location of the data should be mentioned in the publication to ensure that the research data can be found.
5. Besides the documentation on the variable level (codebook), a documentation on the level of the study and the data objects (files, versions etc.) is necessary. This can be achieved through a README file in the repository that provides an overview of the archived data and provides notes for reproduction if applicable (e.g., which software is required? In which order should the scripts be executed? How should the dataset be cited?).
6. It is recommended to also share the instruments, software, and materials in addition to the data.[27]

26 Datenmanagement und Data Sharing in der Psychologie. Einführung und Manual. Herausgegeben vom Forschungsdatenzentrum PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID) [Data management and data sharing in psychology. Introduction and manual], Trier, Germany, May 15, 2013.

27 Sharing commercially available tests always requires the approval of the authors or the respective publisher.

Appendix B:

Exemplary informed consent (if only anonymized data are collected)

Voluntariness:

Participation in the study is voluntary. You may revoke your consent to participate at any time and without stating reasons. This will not cause you any disadvantages. Even if you quit the study prematurely, you are entitled to the compensation for your participation up to that point (i.e., the respective share of money or course credit). You can revoke your consent to store your data until the end of data collection. This will not cause you any disadvantages.

Protection of data privacy:

As no personal data are collected, your personal data cannot be identified after the data collection has been completed, i.e. the dataset is anonymous. Accordingly, a selective deletion of your data record will not be possible after data collection is finished as we will be unable to identify them.

Usage of anonymized data

The results and data from this study will be used for a scientific publication. Anonymity of the participants will be ensured in this process, that is data cannot be related to specific persons. The fully anonymized data from this study will be made accessible as open data on the internet via the data archive _____

Thus, the study follows the recommendations of the German Science Foundation (DFG) and the German Psychological Society (DGPs) for ensuring quality standards in research.

I hereby affirm that I have read and understood the above participant information and agree with the conditions of participation.

Appendix C:

Excerpt from the guidelines of the ethics committee (in press).

Information, debriefing, consent, protection of data privacy law

1. Participant information

   Does the written participant information provided here correspond to the final version?

   Does the informed consent contain an addendum that requires the explicit consent from the participant if the researcher(s) plan to contact them again/continue with the data collection at a later point in time?

   If yes, does it describe how the protection of personal data is ensured in this case?

   In case that minors participate, is the participant information sheet appropriate for this age group?

2. Protection of data privacy

   Has the section about data protection been integrated and highlighted in the participant information?

   Is the data protection section sufficiently detailed and comprehensible for a layperson?

3. Pseudonymization

   Is the way in which the pseudonymization is described comprehensible and does it ensure data protection in accordance with the legal requirements (German data protection law (Bundesdatenschutzgesetz) §3, paragraph 6)?

Final check

Does the participant information meet all requirements?
Does the consent form meet all requirements?
Will there be video and/or audio recordings? If yes, this requires separate consent forms.
Are participant information and consent forms available as separate sheets?
Are potentially relevant additional information from the participants needed (e.g., for EEG, MRT or TMS studies)?
In case minors or other vulnerable populations studied, were any special protective measures undertaken?
In case of online studies, how will the general ethical principles be adhered to (esp. ensuring the voluntariness of participation, anonymity, protection of data privacy)?

Further details can be found in the ethics guidelines of the DGPs www.dgps.de/dgps/aufgaben/ethikrl2004.pdf (3a-e; 6;9) as well as in the DGPs ethics committee proposal guidelines (section C): www.dgps.de/kommissionen/ethik/hinweise zur antragstellung.pdf

Appendix D:

Excerpt from the proposal form of the ethics committee of the Department of Psychology and Education at the LMU Munich (March 14th, 2016)[28]

4) Information about data protection

4.1 What personal data, such as, name, email address, place of residence, or other personal information, are collected?

4.2 Will there be video or audio recordings or other recordings of behavior?

4.3. How will the anonymization or pseudonymization of the collected data be ensured?

4.4a When will the stored data be deleted?

Note for applicants: The personal data (e.g., names, e-mail addresses or other personal data) must be deleted once they are not needed anymore to recruit participants or to contact participants for follow-up questions. It is advised to add a corresponding section to the data protection statement and the consent form. For example: “The deletion of your personal data will occur in accordance with the principles for human subjects research of the German Science Foundation (DFG). Personal data will be deleted once they are not needed anymore to recruit participants or to contact them for follow-up questions”.

Please note: The deletion of personal data has to be documented and verified upon request. In contrast, fully anonymized raw data do not have to be deleted and should be transferred to publicly accessible repositories/databases in accordance with the DFG “Guidelines on the Handling of Research Data”. This is the only way that replicability of the results for further research is assured.

See http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf, section 2 and 3. Participants should be informed about the possible publication of the fully anonymized data. See the “Exemplar of participant information for data protection in case of open data” under “ Appendix 8”.

4.4b It must be stated in the proposal how the pseudonymization or anonymization and the type and time of the deletion of personal data is handled and who is in charge of this.

4.5 Can participants request the deletion of their data at any time?

28 http://www.fak11.lmu.de/forschung/ethikkommission/

Appendix E: Data sharing types and secondary use of data illustrated with an example

The present specification of the DFG guidelines shall be further illustrated using the following fictitious example that combines all of the scenarios of data sharing and secondary use addressed in this document. In reality, all of these scenarios will rarely appear together in the same project. The example is meant as a condensed illustration.

A large research project is planning a longitudinal study over 10 years in which a multitude of variables will be measured. The proposal submitted to the DFG already contains a section about data management and personnel costs for the adequate documentation and sharing of data have been included. Before they are shared, the data will be anonymized in accordance with the HIPAA Safe Harbor Standard and the guidelines by Hrynaszkiewicz et al. (2010) (http://www.bmj.com/content/340/bmj.c181). This means that, for instance, names, e-mail addresses, photos or specific geographic information will be removed.

Two years after the project has launched, data from the first wave of data collection were processed and participating researcher A publishes a paper based on three variables (life satisfaction, neuroticism, and number of friends). The primary data for these three variables are shared alongside the paper (Data Sharing Type 1). In addition, the author of the paper refers to a publicly accessible list of other variables in the dataset.

- Researcher V reanalyzes these three variables and, byusing a modern nonparametric analysis procedure, discovers that the mean effect is even noticeably larger than the one reported in the original publication. He informs A about these results, reports the reanalysis in a blog post and on PubPeer and cites the primary data and the original publication in both cases. Researcher A links to the reanalysis in the repository.

Researcher B, who is also involved in the original project, publishes another paper in which he reports a surprising finding from an exploratory analysis of data from the first wave (based on other variables). He makes the data available according to Data Sharing Type 1.

- Researcher W reanalyzes the data and finds an error in the analysis. She informs researcher B about her discovery; he confirms the error. B and W decide to send a commentary to the journal which results in the publication of a corrected version of the paper.

Researcher C publishes another paper including data from wave 1 of the longitudinal project.

- Researcher X can use these primary data for a meta-analysis (as it is a within-subjects design, the correct standard error of the effect size estimate could only be determined with the primary data). The dataset is cited in the publication; coauthorship is not offered as this is not common practice for meta-analyses.
- Researcher Y can use the primary data to calibrate his priors for a Bayesian analysis (of a different dataset) as he can refer to the known distribution of the existing study. He cites the dataset, but does not offer coauthorship.
- Researcher Z exploratorily tests a new research question with the shared dataset and wants to report this analysis as “Study 1” (in Study 2 the hypothesis will be preregistered and tested confirmatorily with a new dataset). She offers coauthorship to C. However, C refuses as he currently cannot make a substantial contribution to the manuscript due to lack of time.
- Researcher Q conducts a reanalysis and arrives at the conclusion that the effect disappears if “Depressive Symptoms” and “Neuroticism” are controlled for. She informs C about this reanalysis and submits a commentary to the journal. C does not consider the analysis meaningful and publishes a reply to the reanalysis. Thus, there is scientific discourse.

Along the same lines, publications are produced with data from consecutive waves. In the third wave autobiographical memories are recorded in addition to the other data and coded using content categories. Participants are told in the informed consent that these texts will not be shared as they might be used to identify specific persons. Participants are asked whether the data may be stored longer than the obligatory 10 years as a means to scientific quality assurance. The researchers obtain participants’ consent to use the autobiographical memories in the current study. Participants are asked whether the data may be used in further research projects (that are briefly described). The researchers name a person who can be contacted in case participants want to withdraw their consent to further use of their data and to request that their personal data are deleted.

Researcher E publishes a paper based on the autobiographical memories. However, due to data protection, the original texts cannot be shared with the paper. Researcher E explains this in a Readme file in the repository. The frequencies of the content categories per Person, however, are shared (as they are anonymous), so that the reported analyses are reproducible.

- Researcher U has an idea how to code the texts about the autobiographical memories using an alternative coding scheme. He asks researcher E about the primary data and guarantees confidentiality in a written statement. E shares the texts from the participants who agreed to secondary use. As researcher E has contributed substantially to the processing of the texts to enable the analyses by researcher U, researcher U offers her coauthorship which she accepts

After the end of the whole project (i.e., 3 waves and 10 years after the project launch) the principal investigators decide to make use of an embargo for 5 years in order to be able to use the unpublished data exclusively for this period. Further papers are published during this period.

The data publication according to type 1 occurs selectively for those primary data from the complete dataset that are necessary for the reproduction of the reported results.

The complete anonymized dataset – of which most has already been used in publications at this point – is shared at the end of the embargo period. It is specified in the repository how to cite the dataset correctly. The dataset now is a public good. The personal data of the participants who did not agree to an extended storage are deleted 10 years after the end of the project. In the following years, numerous papers based on the dataset are published by other researchers which leads to a large number of citations for the data sharers. In some cases, they are also offered coauthorship. In any case, the data sharers have more citations and more coauthorships than they would have in a “closed science”.