Data Management and Data Sharing in Psychological Science: Revision of the DGPs Recommendations

*German Psychological Society (*Deutsche Gesellschaft für Psychologie - DGPs*) Commission "Open Science"*

(approved by the DGPs Executive Board on June 26, 2020)

Executive Summary

Providing access to research data collected as part of scientific publications and publicly funded research projects is now regarded as a central aspect of an open and transparent scientific practice and is increasingly being called for by funding institutions and scientific journals. To this end, researchers should strive to comply with the so-called FAIR principles (of scientific data management), that is, research data should be findable, accessible, interoperable, and reusable. Systematic data management supports these goals and, at the same time, makes it possible to achieve them efficiently. With these revised recommendations on data management and data sharing, which also draw on feedback from a 2018 survey of its members, the German Psychological Society (*Deutsche Gesellschaft für Psychologie*; DGPs) specifies important basic principles of data management in psychology.

Initially, based on discipline-specific definitions of raw data, primary data, secondary data, and metadata, we provide recommendations on the degree of data processing necessary when publishing data. We then discuss data protection as well as aspects of copyright and data usage before defining the qualitative requirements for trustworthy research data repositories. This is followed by a detailed discussion of pragmatic aspects of data sharing, such as the differences between Type 1 and Type 2 data publications, restrictions on use (embargo period), the definition of "scientific use" by secondary users of shared data, and recommendations on how to resolve potential disputes.

Particularly noteworthy is the new recommendation of distinct "access categories" for data, each with different requirements in terms of data protection or research ethics. These range from completely open data without usage restrictions ("access category 0") to data shared under a set of standardized conditions (e.g., reuse restricted to scientific purposes; "access category 1"), individualized usage agreements ("access category 2"), and secure data access under strictly controlled conditions (e.g., in a research data center; "access category 3"). The practical implementation of this important innovation, however, will require data repositories to provide the necessary technical functionalities.

In summary, the revised recommendations aim to present pragmatic guidelines for researchers to handle psychological research data in an open and transparent manner, while addressing structural challenges to data sharing solutions that are beneficial for all involved parties.

Research data management has become a rather complex task in the face of increasingly demanding data protection regulations on the one hand and rising expectations in terms of the accessibility, findability, and reusability of data on the other—an important task that is not to be underestimated for the quality and relevance of scientific research. Today, the scientific community largely agrees that research data must fundamentally fulfill four criteria (the FAIR principles[1]; see Wilkinson et al., 2016): The data must be "**F**indable," "**A**ccessible," "**I**nteroperable" (i.e., can be integrated with other data and used by as many analytical and visualization applications as possible), and "**R**eusable." These principles are not only included in the guidelines of the German Research Foundation (DFG) for ensuring good scientific practice[2] and in the professional ethics guidelines of the DGPs and the BDP[3] (refer specifically to Section 7.3: "Principles for Research and Publication," Paragraph 14), they are also required by many other funding institutions (such as the European Research Council[4]) for the allocation of funds and are increasingly a prerequisite for the publication of scientific articles in professional journals. At the same time, the European Union's General Data Protection Regulation and its resultant national data protection laws set clear limits to the unrestricted availability and subsequent use of research data.

However, the perspective and interests of all researchers who make their data available and invest the necessary time and effort (we will refer to them as "data sharers" in the following[5]) must also be acknowledged and taken into account. Ideally, data sharing and secondary data use is a "win-win situation" that benefits the entire scientific community. Hence, it is essential that the scientific community acknowledges research data sharing as an important scientific contribution and develops specific incentive systems.

Recognizing this complexity, the board of the German Psychological Society (Deutschen Gesellschaft für Psychologie, DGPs) issued recommendations for the management of research data in psychology in 2016[6]. In the preamble of this document it was stipulated that these recommendations would be evaluated after five years and revised, if necessary (Schönbrodt et al., 2017). The DGPs Commission "Open Science" was charged with this task in the fall of 2018[7]. The revised version of the data management recommendations we are presenting here includes informal feedback from DGPs members and the results of a systematic evaluation of these recommendations (Abele-Brehm et al., 2019; Gollwitzer et al., 2018) as well as the current commentary and discourse in Germany (cf. RatSWD, 2018) and other countries (e.g., Sim et al., 2020).

Further, when considering technical solutions, this commission felt that it was important to focus its considerations closely on what is technically feasible (and equally to base technical solutions on "best

[1] https://www.go-fair.org/fair-principles/
[2] https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/good_scientific_practice/code_gwp.pdf
[3] https://www.dgps.de/index.php?id=85#c2001839
[4] https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf
[5] The term "data sharer" is used here for all those who make their primary data available for secondary use, although we are well aware that such "sharing" is usually preceded by considerable effort (including the planning and execution of a study, preparing the data, and compiling the metadata), which is not really reflected by the term "data sharer."
[6] https://www.dgps.de/fileadmin/documents/Empfehlungen/Datenmanagement_deu_9.11.16.pdf (German) and https://www.dgps.de/fileadmin/documents/Empfehlungen/Data_Management_eng_9.11.16.pdf (English)
[7] In addition to the authors of the original recommendations (Andrea Abele-Brehm, Mario Gollwitzer, Felix Schönbrodt), this commission includes Christian Fiebach, Anne Scheel, Ulf Steinberg, and—as guests—staff members of the Leibniz Center for Psychology Information (ZPID) and representatives of the "Open Science" working group of the German Psychology Consortium Conference (Psychologie Fachschaften Konferenz, PsyFaKo).

practices"). The current recommendations have thus been developed in close cooperation with the Leibniz Center for Psychology Information (ZPID) at the University of Trier. ZPID has been operating a platform for archiving and sharing research data (formerly known as "PsychData", henceforth "PsychArchives"[8]) since 2002. Due to its compatibility with the requirements of the European Union's General Data Protection Regulation as well as the national data protection laws and its clear discipline-specific focus on psychology, PsychArchives was used to verify and ensure the technical feasibility of the recommendations formulated here.

1. Revision of the Data Management Recommendations: Fundamentals

The aim of the DGPs' recommendations on data management remains the development of practical solutions that are effective, efficient, sustainable, and acceptable in everyday research, while taking into account the interests of those who collect and share data, complying with current data protection requirements, and meeting the current standards of openness and transparency. In our view, openness and transparency ("Open Science") are not an end in themselves, but are essential for the purpose of quality management (i.e., preventing mistakes and erroneous conclusions) and for increasing the efficiency of scientific processes. The present revision amends, updates, and refines the 2016 data management recommendations for the following areas:

1. The definition of the terms “raw data,” “primary data,” and “secondary data” (Section 2);
2. The legal considerations associated with sharing data, including data protection, copyright, and right of use (Section 3);
3. Requirements for eligible repositories (Section 4);
4. The appropriate time point for sharing data (Section 5);
5. Third party access and usage rights (Section 6); and
6. Structural challenges and incentives and the management of disputes (Section 7).

Many of the DGPs members who participated in our 2018 survey on the visibility, acceptance, and usefulness of our recommendations (Abele-Brehm et al., 2019; Gollwitzer et al., 2018) were particularly interested in further clarification on the following topics: rights of data sharers, regulations regarding co-authorship and disputes, data protection and copyrights, as well as the options for restricted data sharing or a definition of the specific conditions for the secondary use of shared data. These requests are addressed here.

2. Definitions of Terms: Primary Data, Metadata, Secondary Data

2.1 Primary Data

Given that the term "primary data" is often referred to in the following, this type of data will be defined first

[8] https://leibniz-psychology.org/angebote/archivieren,/

and then differentiated from the term “raw data”[9]. In addition, the distinction between primary and secondary data is common and will be discussed below (see Section 2.3). *Raw data* are the original recordings collected from a source, such as the marks on a paper questionnaire, drawings, audio or video recordings, eye movement measurements, or neurophysiological or peripheral physiological recordings (e.g., EEG, heart rate). Raw data can therefore be defined as the first, "non-transient" form of data. In many cases, but not necessarily, raw data are already available in digitized form. We define *primary data* in psychological research as the first transfer of raw data into a digital format, e.g., the code "1" for a yes response in a questionnaire. Raw data and primary data are often equivalent, for example, when the respondents' answers were collected by means of experimental software or in an online survey and then immediately stored in a digital format. Thus, primary data in psychology are unaltered (i.e., untransformed, not aggregated, etc.) quantitative or qualitative data available in digital form, e.g.:

- Each manipulated and measured variable of every experimental session of every study participant in an experiment;
- Each response of every person to each item in a survey;
- Original wording of inputs in free text fields;
- Digitized video recordings;
- Downloads or screenshots of social media content (e.g., Facebook profiles or Twitter messages);
- Transformed (neuro)physiological data (such as EEG or fMRI 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 the data of cases that were excluded from the analyses (with the exception of those cases in which participants withdrew their consent during or after data collection). To summarize, we define primary data as the set of all data points collected during the course of a study or project, as initially digitized, but otherwise in a completely unaltered form.

Primary data should be publicly available in an open and freely accessible data format (to fulfill the FAIR principles of accessibility and interoperability). Data should be made available in a standardized file structure, as this also facilitates its secondary use. As a standardized file storage structure for MRI data, the Brain Imaging Data Structure[11] represents an example of "best practice."

[9] This distinction is often based on the specific discipline. The National Research Data Infrastructure (NFDI), for example, makes a more general distinction between "primary research data" and "research data," with the former being defined as "acquired raw data that have not been processed in any way, commented on, or tagged with metadata." In this sense, research data include all data, secondary analyses, visualizations, results, etc., that are generated throughout the course of the research process (https://www.forschungsdaten.info/praxis-kompakt/glossar/#c269824). The DFG defines primary research data as "data that has been acquired in the course of scientific research, experiments, measurements, surveys, or polls. These data form the basis for scientific publications" (https://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf). However, discipline-specific definitions are also mentioned here.

[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. For EEG and MEG data see, e.g., the blog "Best Practices in Data Analysis & Sharing in Neuroimaging using MEEG" (https://cobidasmeeg.wordpress.com/).

[11] https://bids.neuroimaging.io; We also refer to the project "psych-DS," which is currently under construction, that aims to develop a file structure for psychological data in general, similar to BIDS, see https://github.com/psych-ds/psych-DS

2.2 Metadata

Primary data require *unique* and *rich metadata* to ensure that they are findable and reusable. Metadata provide structured information on the primary data set, including the information necessary to verify and understand both the results and conclusions of the study for which the primary data were collected and to evaluate the reuse potential of the primary data set. This includes not only a "data dictionary" (i.e., clear labeling and descriptions of the variables in the data set), but also information on the research sample (e.g., recruitment, field access, sampling procedure), the process of data collection, data processing, and data analysis, and the data generation process (i.e., how the data were generated and by whom; RatSWD, 2018).

Metadata may have a content-related or technical-administrative function. Technical metadata encompasses the information pertaining to all relevant parameters of data collection (in the case of EEG data, e.g., the sampling rate and filtering). This information is necessary to ensure both the comparability of empirical replications and the possibility of secondary data analysis (according to the FAIR principle of "reusability"). This information can be provided as notes in the raw data sets or as accompanying documentation (i.e., as a text file) in the repository. To achieve the aim of interoperability, universal metadata standards have been established, such as the "Data Documentation Initiative"[12] or the "Dublin Core Metadata Initiative"[13]. Discipline-specific information on metadata in psychology is available, for example, via DataWiz, a data management system developed by ZPID specifically for use in psychology research.[14]

2.3 Secondary Data

Secondary data refers to data that has undergone some type of initial processing[15]. This includes, for example, the transformation of variables, aggregation of individual observations, calculation of scale values, averaging of event-related potentials, or exclusion of values that are assumed to affect the robustness of the analytical results (e.g., extreme and outlier values, invalid cases). In the interest of reproducibility and reusability of data, primary data should always be provided; however, legal or ethical reasons may require that secondary data derived from primary data be provided (see Sections 3.2 and 3.3). In such a case, the program code should be provided to allow the user to trace how the secondary data set was generated from the primary data.

3. Legal Aspects of Data Sharing

3.1 Data Protection

Legal data protection restrictions must be taken into account in the planning phase of any research project and are essential for later data sharing. The relevant sections of the Federal Data Protection Act (Bundesdatenschutzgesetzes, BDSG[16]) and the data protection laws of the federal states must be observed. In the BDSG, sections 27 (Data processing for purposes of scientific or historical research and for statistical purposes), 46 (Definitions), 47 (General principles for processing personal data), 50 (Processing for

[12] https://ddialliance.org

[13] https://dublincore.org

[14] https://lifp.de/psychologische-metadaten

[15] Other authors define "secondary data" differently, for example, as data that is reused for a purpose other than that for which it was originally collected (see Hox & Boeije, 2005).

[16] http://www.gesetze-im-internet.de/bdsg_2018/ - English version: https://www.gesetze-im-internet.de/englisch_bdsg/

archiving, scientific and statistical purposes), 51 (Consent), and 64 (Requirements for the security of data processing) are of particular importance. Further information can also be found in specific publications (e.g., RatSWD, 2020).

3.2 Personal Data

Personal data (i.e., "any information relating to an identified or identifiable natural person"; Section 46 (1) of the BDSG) must be anonymized or pseudonymized[17]. This is to ensure that persons cannot be identified based on any combination of different characteristics that have been recorded—even data that have been acquired in different studies with the same participants—(e.g., Section 47 of the BDSG). However, legal requirements for data protection apply not only at the level of individual persons, but also at relevant aggregate levels: In the case of sensitive issues (e.g., illegal behavior, suicide rates), the extent to which individual schools, companies, etc. can be clearly identified in the data or identified after merging data sets must be taken into account (RatSWD, 2020).

Before the start of any investigation, documentation is generally required detailing the basis for and manner of personal data collection and processing as well as the technical and organizational measures that will be taken to ensure data protection and data security. Furthermore, when a high risk to the rights and freedoms of natural persons is to be expected as a result of data collection or processing, a data protection impact assessment is required (e.g., Martin et al., 2020).[18] This is particularly important if data of a highly sensitive nature are collected (e.g., ethnic origin, political, religious, or philosophical beliefs, details about sex life, as well as physiological or GPS data from tracking devices, which allow conclusions to be drawn about a person's health status or very easily facilitate reidentification).

Data sharers are required to document the anonymization/pseudonymization of the data. A list documenting these processing activities must be provided to the supervisory authorities upon request; in the case of high-risk data (see above), the impact assessment must also be documented. When this information is available from the data sharers, legal action can no longer be taken against them in the event that secondary data users violate data protection rights in the course of secondary use.

3.3 Consent

It is crucial, not only from a legal point of view (RatSWD, 2020), but also from the perspective of research ethics (DGPs, 2018), that participants are thoroughly informed about the benefits, the risks, and the types of data collection, as well as about the purposes of data use, data storage, and further data utilization in a transparent and understandable way so that, fully informed, they can freely consent to the processing of their data. The voluntary nature of participation must always be guaranteed. At all times, participants should have the option of terminating their participation in the study. Subjects should be guaranteed the opportunity to view and correct the recorded data (unless the data have been anonymized). Consent is mandatory for video

[17] Pseudonymization means "the processing of personal data in such a manner that the data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data cannot be attributed to an identified or identifiable natural person" (Section 46 (5) BDSG).

[18] http://publica.fraunhofer.de/documents/N-586394.html

and audio recordings and must be obtained accordingly; the process of obtaining consent must be documented (see DGPs, 2018).[19]

The German Ethics Council distinguishes between different types of informed consent (see RatSWD, 2020; p. 27). By giving *blanket consent,* study participants agree to an indefinite future use and disclosure of their data. When *dynamic consent* is given*,* study participants are repeatedly contacted to obtain their informed consent to specific research questions. They are in regular contact with a database (e.g., via an online platform or telephone contact). *Cascading consent or meta consent* is an extension of dynamic consent. In this case, consent does not necessarily have to be obtained separately for each new research question, and research participants can choose between different options.

Recital 33 of the General Data Protection Regulation (GDPR) provides the possibility of "broad" consent specifically for scientific research. The European legislator assumes that the purpose of processing personal data for scientific research cannot always be fully specified at the time when personal data are collected. Therefore, it authorizes for study participants to give "broad" consent (a) for certain areas of scientific research or parts of research projects, (b) to the extent permitted by the purpose pursued, and (c) in accordance with the recognized ethical standards in scientific research. These "certain areas," however, must be related to the original research goal.

Apart from their basic consent, study participants must be informed that their anonymized data may be made available for secondary use by third parties and that the purpose, nature, and scope of this secondary use is currently not foreseeable. Explicit informed consent for secondary data use must be obtained when data cannot be fully anonymized (Metschke & Wellbrock, 2002). When data are fully anonymized, consent is not legally required (since individual identification is no longer possible), but it is nevertheless mandatory from a research ethics perspective (DGPs, 2018).[20]

Not fully anonymized data of individual participants who have refused to give their consent for potential secondary use may not be shared. An appropriate explanation should be provided if data cannot be shared, (e.g., in a footnote in the publication and in accompanying documentation of the data set in a repository). However, 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 or pseudonymized partial data can be shared.

To guarantee participants the right to view their own data or to withdraw their consent retroactively after data collection, a record may be kept for a limited time to allow the identification of participants using pseudonymization keys. This unblinding list, which contains the pseudonymization keys that are assigned to real names, is deleted after a predetermined time period (e.g., X hours/days after the end of data recording) that is clearly communicated to the participants.

3.4 Copyright and Rights of Use

While primary data are generally not protected by copyright (Guibault & Wiebe, 2013; Hillegeist, 2012;

[19] Templates (in German) for declarations of consent are provided by the ethics committee of the DGPs on the TransMIT Centre for Scientific and Psychological Services (ZwpD) website: https://zwpd.transmit.de/zwpd-dienstleistungen/zwpd-ethikkommission/vorlagen-antragstellung

[20] This may include data collected in earlier studies.

Spindler & Hillegeist, 2011), questions pertaining to rights of use must always be clarified before publishing these data. When data are collected by the scientific staff of a research institution, for example, the rights of use are usually held by that institution[21]. However, the legal status of the employer-employee relationship in terms of research data is very complex, making it difficult to make general statements regarding this issue[22]. Therefore, questions on rights of use must always be resolved before research data is published.

4. Requirements of a Suitable Repository

The primary data should be made available in digital form in a trustworthy repository. Important quality features of trustworthy repositories are:

- the economic autonomy as well as the scientific professionalism of the institutional provider;
- the accessibility of data (it must be possible to obtain the data free of charge and according to a fixed, graduated access category model; see Section 6.2);
- the citability of data (a "Digital Object Identifier" [DOI], referencing a unique version of the data, must be assigned);
- the observance of data protection according to EU regulations in the country where the repository stores the data;
- the clarification of data property rights (storing data must not involve ceding exclusive rights of use to third parties[23]); and
- the persistence of data (i.e., clarification must be provided as to how long and under what conditions the data will continue to be available without restriction in the event of the dissolution of the repository or its operating institution).

For these reasons, trustworthy open repositories (e.g., PsychArchives at ZPID[24], datorium at GESIS[25], or a well-developed university repository) are preferred over journal repositories. We strongly advise against storing data on private or personal university websites. In addition, the institution providing the repository service should be able to provide qualified information and advice on primary data storage. When choosing a repository, constraints imposed by legal regulations or research ethics guidelines (e.g., prohibition of data storage on servers located in foreign or non-European countries) need to be considered.

[21] Under certain conditions this may be different and the rights of use are held by the researchers themselves; see Hillegeist (2012) and https://blogs.hrz.tu-freiberg.de/oersax/urheberrecht-des-wissenschaftlichen-personals-an-hochschulen

[22] Claims to research data may be governed by, for example, copyright law, ancillary copyright, patent law, civil law (labor law), and constitutional law (scientific freedom). Hillegeist (2012) offer an overview, see https://www.forschungsdaten.info/themen/rechte-und-pflichten/urheberrecht/ and https://www.forschung-und-lehre.de/wem-gehoeren-forschungsdaten-1013/.
Therefore, it is advisable to contact the university administration and, where necessary, aim to establish a general regulation for the handling of research data that is committed to open science principles and allows and promotes the publication of data collected and compiled by employees of the institution. Such a research data guideline will eliminate ambiguities and the need for researchers to make enquiries in each specific case. Examples of such research data guidelines are provided by the Technical University of Munich (TUM; https://www.it.tum.de/projekte/forschungsdaten-management) and Heinrich Heine University Düsseldorf (HHU; https://www.uni-duesseldorf.de/redaktion/fileadmin/redaktion/Fakultaeten/Medizinische_Fakultaet/MedRSD/Dok_GWP/Forschungsdaten_Richtlinie_2015.pdf).

[23] The *simple* rights of use, i.e., the right to archive and reproduce, must be conferred to the operator of the repository so that repository services can be provided in the first place.

[24] https://psycharchives.org

[25] https://data.gesis.org/sharing/#!Home

5. Timing and Type of Data Sharing

With regard to the time of data publication, we distinguish—as in the original recommendations (Schönbrodt et al., 2017)—between two types of data sharing.

5.1 Sharing Data as Part of a Publication ("Type 1")

With the publication of a manuscript, the person or group who collected the data (the data sharers) should provide all primary data and associated metadata necessary to reproduce the published results, regardless of the context in which they were produced (e.g., a third-party funded research project or a student thesis). The term "publication" explicitly includes manuscripts that have already been made available to the public as so-called "preprints" prior to publication in a scientific journal. For manuscripts published according to standard procedures—i.e., papers that become publicly available only once they are published in a journal—providing primary data and metadata to reviewers during the review process is required by an increasing number of journals. Unless otherwise specified, the first author or the corresponding author is responsible for taking these steps and providing proper documentation of the data.

For simulation studies, the data-generating code should be shared. If the simulated data cannot be adequately reproduced by the code, or if reproducing the simulated data is excessively costly or laborious, it may be necessary to share the simulated data files as well. Documentation of the program code must include all information necessary to reproduce the data.

Generally, when documenting the primary data, variables that were assessed within the framework of the study but not included in the corresponding publication should also be reported (see "standard reviewer disclosure request"[26]). Data from these variables are shared in the primary data set when they are later used for a publication or when the research project is concluded and the complete data set is made available (Data Sharing Type 2; see Section 5.2).

Prior to publication—and ideally even before data collection—it is necessary to clarify whether all data or only selected data can be provided, and whether— e.g., due to research ethics reasons—only secondary data can be shared. These considerations must be documented (see Sections 3.2 and 3.3).

5.2 Sharing Data after Project Completion (“Type 2”)

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"[27]. This includes all relevant data of the project that are not yet part of a publication, including the accompanying metadata (see Section 2.2).

It is at the discretion of the person responsible for the project to decide which data are "relevant." Examples of irrelevant data might be data based on flawed code or collected in highly exploratory pilot studies. To

[26] https://osf.io/hadz3/
[27] www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

counter the problem of publication bias, all data that have produced results that do not conform to expectations should also be provided (except in cases where technical or manual error has occurred). We consider it logical in this context to require project managers to explicitly state which studies have been conducted and where the corresponding primary data are located (as well as how they can be accessed and reused) in the final report of a project. Reviewers should also assess final reports in terms of whether the documentation of data sharing is sufficiently transparent and accurate.

The time at which a project is considered completed depends on the complexity of the project. However, in projects funded by third parties, the submission of the final report usually constitutes the end of the project. Many funding institutions require that the data collected in the project be shared as soon as possible after the project has been completed. Given this background, it should be clear to all project leaders that research data management is a task that must be considered from the start (and not once the project has been completed). Yet there may be cases where data preparation and data sharing is only possible when a project is nearing completion, and data sharing is delayed accordingly. In this case, it is recommended to set up the appropriate data structures in the repository, including the corresponding metadata, and to include a text file indicating the expected delivery date (see also Section 5.3).

Data Sharing Type 2 applies particularly to projects for which both the scope and completion are properly defined (as is typically the case in third-party funded research projects). For continuously running projects (e.g., financed by university funds), the exact time of project completion may be difficult to define. Here, too, data should be shared in a timely manner.

5.3 Embargo on Use

Data sharers can define an embargo for secondary use. This means that data collected in the course of a research project that have not yet been used for publications will be stored in a repository immediately after project completion—as explained in Section 5.2—but will not be available to third parties until a later date. An embargo on use may be justified, for example, to allow the temporary safeguarding of intellectual property and/or to provide additional protection for the data sharers from adverse consequences. An embargo period for Type 2 data sharing should typically not exceed five years following project completion. In general, the reasons for implementing an embargo must be stated.

Typically, data published as part of a publication (Data Sharing Type 1) should not be subject to an embargo. In exceptional cases (e.g., extremely complex data collection), researchers may impose an embargo on these data; here, regulations stipulated by the journal in which the article was published must also be taken into account. However, an embargo in connection with Type 1 data sharing should be significantly shorter than the embargo period recommended for Type 2 data sharing of a maximum of five years. When an embargo is imposed on Type 1 data sharing, measures must be taken to ensure that, upon publication, the data are available by request to reproduce the reported results (see example 3 in Section 6.2).

A suitable method to document an embargo on data use includes reporting information on the type of embargo, the justification, the length of the embargo, as well as other details about the data not yet provided (metadata; referring to a codebook or data dictionary, if applicable; see Section 2.2) in a standard text file stored in the repository. An embargo can also be implemented in specific repositories by initially (and temporarily) assigning a more restrictive access category (see Section 6.2), with the dataset automatically changing to a less restrictive access category after a specified time period. This means that after the embargo expires, the data will be made available for subsequent use under the previously defined conditions. As a "best practice" approach to imposing an embargo, we recommend a "file only embargo," under which the actual data files are restricted, but the corresponding metadata are already openly accessible (and thus discoverable) at an earlier stage. Furthermore, the lifting of the embargo should be automated as a function of the repository, requiring no further action by the data sharer.

5.4 Preventing Unnecessary Duplication of Data Sets

Secondary data (i.e., any transformation or selection of data derived from a primary data set; see Section 2) should never be submitted as a new primary data set—not only would this be inefficient, it could also lead to serious distortions in further applications (e.g., if the two data sets were later included as independent data sets in a meta-analysis). Rather, secondary users should always reference the primary data set originally released by the data sharers and clearly document the process of how the derived data sets were created from the original data.

However, the distinction between Type 1 and Type 2 data sharing described in Sections 5.1. and 5.2. allows for a gradual release of (partial) data sets, possibly with overlapping content. To reconcile this with the need to prevent duplication, data sharers could publish the entire dataset (as per Type 2) under an embargo (see Section 5.3) and, by means of a reproducible script, generate a partial data set that is immediately made available as a Type 1 release in conjunction with a publication. The repository should clearly indicate that this data set has been derived from a separate primary data set. Furthermore, the repository should present the primary data set and all derived partial data sets bundled together so that the interconnection between them is visible. As a general rule, the number of partial data sets for each data collection should be kept as low as possible, and the relationship between these "related" partial and primary data sets must be disclosed and communicated in a transparent manner.

6. Licenses and Access Categories

6.1 General Considerations

Researchers who produce primary data ("data sharers") should have the exclusive right of first use of this data (even if this right is not legally enforceable in the strict sense). The same applies in cases of Type 2 data sharing (see Section 5.2) where, at the time of sharing, the data have not yet been used for their own purposes by the researchers who shared the data. Data sharers can safeguard their right of first use by means of a temporary embargo (see Section 5.3).

Researchers who use data shared by others have to cite the data appropriately[28]. For this purpose, the data in the repository should be accompanied by a citation reference (including a persistent identifier; DOI - see Section 4). Secondary users are required to analyze the data in such a way that does not infringe the rights of the participants in the original study. This is the responsibility of the secondary users, and they must strictly adhere to the specific terms of use stipulated by the data sharers.

[28] https://www.force11.org/datacitation

The secondary use of data has to meet the same requirements of transparency and scientific diligence as the primary use. The scientific standards for reinterpretation of the data have to be those that are valid at the time of the secondary analysis. Conversely, when evaluating original analyses in the context of a reanalysis, it is only fair to apply the scientific standards that were valid at the time of the original analyses.

All the provided data files (including any copies that were made) must be permanently deleted by secondary users after the expiration of the contractually stipulated right of use.

6.2 Access Categories

In accordance with data protection laws, aspects of research ethics (see DGPs, 2018), or legitimate scientific interests of the data sharers, it may be necessary to restrict access to or the use of a data set, regardless of whether the shared data are Type 1 or Type 2 (see Sections 5.1 and 5.2). This should always be done by carefully considering issues of data protection and research ethics in a risk analysis, taking into account both the risk of misuse and the severity of harm in the event of misuse. For example, both a high risk of participant reidentification paired with low expected harm and a low risk of reidentification paired with high expected harm may require a high level of access restriction. We therefore differentiate four access categories, which are characterized in the following as 0, 1, 2, and 3. Each category combines a technical access restriction (availability) with a usage rights license to regulate the authorized use of the respective data sets. The term “license” here refers to a contractual agreement by which data sharers grant a second party permission to commit acts that would violate the rights of the data sharer without this explicit agreement[29]. Please note, however, that questions pertaining to copyright and rights of use must be clarified before a data set can be licensed (see Section 3.4).

- **Access category 0 ("open data"):** There are no restrictions of any kind on access or use of the data. The data use can—depending on the respective platform or the functionalities of the repository—be digitally traced (“tracking”). In some cases, secondary data users may be requested to provide information on their identity, affiliation, and/or intended secondary use[30]. Legitimate standard licenses, available, e.g., from the Creative Commons[31] or the Open Data Commons[32], are commonly used for this purpose.
- **Access category 1 ("open data/conditional access"):** Access or secondary use is subject to certain conditions stipulated by the data sharers to which the secondary data users must explicitly agree. In these cases, license agreements with standard terms and conditions (see below for examples) are used. Access is only granted after the secondary users have agreed to these contractually stipulated conditions (without obtaining prior consent of the data sharer). Any breach of the conditions is subject to legal recourse by the data sharers (see Section 7). Some repositories may collect information about the secondary users (identity and affiliation) and/or the purpose of the secondary use and provide data sharers with this information.
- **Access category 2 ("restricted access"):** In this category, access or secondary use is contingent

[29] https://www.forschungsdaten.info/themen/rechte-und-pflichten/urheberrecht
[30] The extent to which this information is transmitted to the data sharer depends on the wishes of the data sharer and the functionality of the repository.
[31] https://creativecommons.org/licenses
[32] https://opendatacommons.org

on secondary users' agreement to further conditions imposed by the data sharers in addition to those applicable in access category 1. These are not "standard cases" (as in access category 1), but individualized contracts between data sharers and secondary users. Access is only granted when both parties to the contract fully agree to the terms and conditions that are defined in the contract. Breaches of contract are subject to legal recourse by the data sharers (see Section 7).

- **Access category 3 ("secure data"):** Data access is only possible in compliance with restrictions mandated by data protection law (e.g., online via a suitably secured channel or in person at a data center). Both the type and extent of use is stipulated by the data sharer and monitored by the respective facility where the data is located.

Selecting the appropriate access category is the responsibility of the data sharer and should comply with all requirements of third-party funders. Reasons for the decision should be provided and documented in the course of data publication. To promote openness and transparency, we follow the principle "as open as possible, as closed as necessary."

Access category 0 is applicable in all cases in which (a) personal data (see Section 3.2) have either not been included in the data set or are irrelevant in terms of the research question (and can be removed from the open data set) and (b) there are no plausible reasons (e.g., research ethics) for limiting or restricting the secondary use from the perspective of the scientific community.

Standard terms and conditions for access are defined in **access category 1.** The key feature of access category 1 is the establishment of a *standard license agreement* for secondary use. Although individual repositories may formulate these standard license agreements differently, the full advantage of this access category for the scientific community only unfolds when a certain degree of homogeneity of the corresponding contractual conditions is achieved: Once general standards have been established, both data sharers and secondary data users will be able to develop efficient work routines and thus avoid spending an undue amount of time dealing with legal matters. Standard terms and conditions of access may include, for example, the following aspects:

- the exclusive use of a data set for scientific purposes ("scientific use"; see Section 6.3),
- the obligation to make reference to one or more relevant sources (e.g., the original publication) in any publication resulting from secondary data use,
- the procurement of any necessary declarations of consent (e.g., from a test publisher who holds the rights to the normative data of a test),
- the explicit contractual declaration that no attempt will be made to reidentify study participants, or
- the obligation to adhere to usage restrictions previously negotiated with the study participants (e.g., to disclose any clinically significant incidental findings revealed in the course of the reanalysis; see DGPs, 2018)

In **access category 2,** data sharers can define individualized restrictions on use in a license agreement. These restrictions must be substantiated in a comprehensible manner. The following examples illustrate the meaning of "comprehensible" in this context.

- **Example 1:** The data provided is the normative data set of an ability or achievement test, and is only able to yield valid results if the correct answers are not publicly available. The data set (which necessarily includes this information) can be reused, but the secondary users are not allowed to publish this information under any circumstances.

- **Example 2:** There is a risk that secondary users may attempt a biased and invalid reanalysis of the original data in their pursuit a particular research agenda. The contractual condition here could stipulate that in the event of a planned publication of the reanalyzed data, the secondary users are always required to give the data sharers the opportunity to review or comment on the reanalysis before submitting the conclusions of the research to a journal, etc. This approach is only recommended in very rare cases and must be justified in a publicly documented, transparent manner by the data sharer.
- **Example 3**: Data sharers consider the possibility of imposing an embargo (see Section 5.3). The supplementary agreement could specify that during the embargo period, the data may only be analyzed for the purpose of reproducing previously published results or used for meta-analyses, but not for any further research. However, such a restriction should only be allowed if the end of the embargo period is clearly stated and the data are subsequently made available to secondary users in a less restrictive access category.

Restricting access as defined in **access category 3** is appropriate when the risk analysis indicates that individuals can be easily identified or when the disclosure of personal information is likely to cause harm to the relevant persons. Such information could include highly sensitive data such as imaging data with identification of cranial shape or facial features, genome data, but also survey data on sensitive topics in clinical psychology or industrial and organizational psychology. Secondary users must be required to sign an appropriate, restrictive confidentiality and nondisclosure agreement. Use of this access category can, in individual cases, also be warranted if there are reasonable grounds to assume that data sets will be used improperly (cf. Lewandowsky & Bishop, 2016).

Provided there are no legal or research ethics objections, a restriction on access can be lifted by the data sharers at any time. Restricting access to the data retrospectively, however, should only be carried out if a higher access category is later deemed necessary based on legal or ethical grounds.

6.3 Scientific Use

Data files (including experimental material, primary data, etc.) that are exclusively restricted to the context of scientific analysis and discourse are called "scientific use files" (SUF). "Scientific use" can be defined either in terms of the group of users or the specific objectives of use (or both). The definition of the *user group* could stipulate, for example, that access is only granted to people who work at a national or international research institution and/or can verify their academic training (typically a PhD). However, we do not recommend defining “scientific use” via the user group, as automatic verification of such information is susceptible to error. Instead, access category 2 restrictions should be stipulated in individualized user contracts if a restriction of the user group is desired and warranted.

In terms of its *intended use,* "scientific use" includes (a) reanalyzing data for scientific quality assurance, (b) presenting illustrations or practical exercises in university teaching, (c) testing the applicability of a specific statistical procedure or evaluating a new method for estimating a statistical parameter, or (d) using existing data to investigate new research questions.

In principle, it is prohibited to share material classified in access category 1 or higher with third parties. When using SUF for teaching purposes (e.g., to illustrate a particular analysis or to test the applicability of a particular model), each student must individually obtain the relevant data from the repository. It is important to note here that, even in the educational context, all students are bound by the conditions defined in the respective license agreements.

SUF sharers should clearly define the conditions for the expiration of secondary users' rights of use: Upon expiration of the rights of use, e.g., when the intended use is accomplished or abandoned, all existing copies of the SUF must be deleted by the secondary user.

7. Structural Challenges and Recommendations

7.1 Acknowledging Data Sharing and Incentive Structures

As we mentioned in the introduction, data sharing and secondary data use is intended to be a "win-win situation" for the entire scientific community, contributing to quality assurance in science and promoting the generation of new, robust findings and methodological developments. This can only be successful, however, if both the sharing of data and its secondary use are undertaken within the spirit of a cooperative and productive maximization of the collective interest. This requires not only the willingness for openness and transparency on the part of the data sharers, but also the willingness of the secondary users to adhere to the established rules of good scientific practice. This calls for the due recognition of the time and effort that is inevitably involved in making data available by those who collected it.

At present, there is intensive debate in various disciplines about the possible forms of such recognition and the structural incentives that can be provided by the academic system[33] that are not—strictly speaking—included in these present recommendations on data sharing. At this point we can only offer suggestions and intellectual impetus. One form of acknowledgment could be the possibility to recognize data sharing—for example, when evaluating scientific CVs—as an important scientific productivity indicator like text-based publications. At present, some journals publish primary data and their corresponding documentation, thus enabling citation of the data[34,35]. When data sharing is recognized as a separate category in CVs, its contribution can also become more visible. Bestowing awards (such as the DGPs Award for Quality Assurance in Psychology or the Leamer-Rosenthal Prizes for Open Social Science sponsored by the Berkeley Initiative for Transparency in the Social Sciences[36]) is also an appropriate, though by no means complete, way to recognize the willingness of young researchers to share their data.

[33] e.g., see https://psyarxiv.com/6btc3

[34] https://openpsychologydata.metajnl.com

[35] https://www.nature.com/sdata

[36] https://www.bitss.org/lr-prizes

Professional societies, funding institutions, scientific journals, as well as local institutions are all called upon to develop and implement suitable incentive structures. These incentive structures must aim to provide commensurate recognition of the willingness to share data and the time and effort that is involved in doing so.

7.2 Co-authorships

If the secondary use of the data results in a publication, the question arises as to whether and under what circumstances should co-authorship be offered to the data sharers. Given that—certainly for the time being—authorship in a published work is one of the most frequently used indicators of scientific productivity, it could be argued that including the data sharer as a co-author in publications resulting from the secondary use of data is the simplest and most appropriate way of recognizing data sharing. The DGPs Commission "Open Science" has extensively discussed this argumentation and made the following recommendation: We are of the opinion that the mere sharing of data does not merit a co-authorship of the original authors, because co-authorship requires a "genuine, identifiable contribution to the content of a research publication of text, data or software" (DFG Guidelines for Safeguarding Good Research Practice, September 2019; Guideline 14; see also Section 7.3, Paragraph 13, Statement b of the Professional Ethics Guidelines of the DGPs and the BDP[37]). Such contributions usually go beyond simply sharing data and entail significant contributions to the manuscript itself. Furthermore, we also believe it to be inappropriate for original authors to only provide their data after they have been offered the prospect of co-authorship of manuscripts resulting from the secondary use of their data.

In many cases, however, it can be useful for secondary users to contact the data sharers, for example, to clarify misunderstandings resulting from the nonreproducibility of an analysis or to discuss whether the data sharers' co-authorship is warranted for works resulting from the reanalysis.

The Contributor Roles Taxonomy (CRediT)[38] is a useful tool for representing the different roles and types of contributions of the different co-authors. Currently, a number of journals require a classification of all authors according to the 14 defined CRediT contribution roles (see also McNutt et al., 2018; Holcombe, 2019).

It remains to be seen how routine data sharing and data reanalysis by secondary users will change the view of traditional productivity indicators (such as the number of publications in an academic curriculum vitae) inherent in science. Yet there is no doubt that such changes are necessary and that data sharing (even in cases resulting in no authored manuscript) must also be viewed as a scientifically productive practice worthy of recognition.

7.3 Procedures for Resolving Disputes

In these recommendations, all references to current legislation (e.g., on data protection and copyright; see Section 3), to guidelines of good scientific practice, or to the professional ethics guidelines of the DGPs or the BDP are binding and in principle subject to sanctions. Violations, for example, of the Federal Data Protection Act, are liable to civil and possibly even criminal prosecution. The DFG's 2019 Guidelines for

[37] https://www.dgps.de/index.php?id=85#c2001838

[38] https://casrai.org/credit

Safeguarding Good Research Practice define how scientific misconduct should be dealt with (see Guidelines 18 and 19), and the professional ethics guidelines of the DGPs or the BDP state that violations against them are liable to prosecution and, if necessary, punishment by the Court of Honor of the DGPs[39] and the BDP. Yet many of the recommendations formulated here are not legally binding. For this reason, it is all the more important to address the main and potentially conflicting issues from the outset in standardized agreements (as envisioned in access category 1) or—if necessary—in specific agreements (as envisioned in access category 2) to ensure appropriate safeguards. Any violations of such agreements constitute scientific misconduct; they can be prosecuted in civil court and—if the involved parties are members of the DGPs—can be brought before the Court of Honor of the DGPs (cf. the Rules of Honor of the DGPs[40]). To deal with disputes between secondary users and data sharers that do not definitively fall into the category of "scientific misconduct," but rather represent different points of view, we recommend the establishment of an ombuds committee (ideally authorized by election) that can be consulted by the involved parties, provided they are members of the DGPs.

Conclusion

Science is changing. Driven by new technological advances, urged forward by the so-called replication crisis, and compelled by new demands from journals and funding agencies, the availability of open research data is rapidly emerging as the standard in psychology research.

The recommendations outlined here are intended to provide a framework that will facilitate this transition to more openness and transparency for the benefit of both data sharers and secondary users, leading ideally to a long-term improvement in the overall quality of psychological research.

References

Abele-Brehm, A., Gollwitzer, M., Steinberg, U. & Schönbrodt, F. (2019). Attitudes towards Open Science and public data sharing: A survey among members of the German Psychological Society. *Social Psychology, 50*, 252-260. https://doi.org/10.1027/1864-9335/a000384

Deutsche Gesellschaft für Psychologie DGPs (Eds.) (2018). *Ethisches Handeln in der psychologischen Forschung [Ethical behavior in psychological research]*. Göttingen: Hogrefe.

Gollwitzer, M., Schönbrodt, F. D., Steinberg, U. & Abele-Brehm (2018). Die Datenmanagement-Empfehlungen der DGPs: Ein Zwischenstand [The data mangement recommendations by the DGPs: An intermediate result]. *Psychologische Rundschau, 69*, 366-373. https://doi.org/10.1026/0033-3042/a000415

Guibault, L. & Wiebe, A. (Eds.) (2013). *Safe to be open*. Göttingen: Universitätsverlag Göttingen.

Hillegeist, T. (2012). *Rechtliche Probleme bei der elektronischen Langzeitarchivierung wissenschaftlicher Primärdaten [Legal problems with the electronic long-term archiving of scientific primary data]*. Göttingen: Universitätsverlag Göttingen.

[39] https://www.dgps.de/index.php?id=78#c281

[40] https://www.dgps.de/fileadmin/documents/Fachgruppen/DGPs_Ehrengerichtsordnung_2017.pdf

Holcombe, A. O. (2019). Contributorship, Not Authorship: Use CRediT to Indicate Who Did What. *Publications*, *7*(3), 48. https://doi.org/10.3390/publications7030048

Hox, J. J. & Boeije, H. R. (2005). Data Collection*,* Primary vs. Secondary. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (Bd. 1, S. 593–599). New York: Elsevier.

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

Martin, N., Friedewald, M., Schierung, I., Mester, B. A. & Hallinan, D. (Eds.) (2020). *Die Datenschutz-Folgenabschätzung nach Art. 35 DSGVO: Ein Handbuch für die Praxis [The data protection impact assessment according to Art. 35 DSGVO: A practice manual]*. Karlsruhe: Fraunhofer Verlag.

McNutt, M. K., Bradford, M., Drazen, J. M., Hanson, B., Howard, B., Jamieson, K. H., Kiermer, V., Marcus, E., Pope B. K., Schekman, R., Swaminathan, S., Stang, P. J. & Verma, I. M. (2018). Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, *115*(11), 2557–2560. https://doi.org/10.1073/pnas.1715374115

Metschke, R. & Wellbrock, R. (2002). *Datenschutz in Wissenschaft und Forschung [Data protection in science and research]*. Berlin: Berliner Beauftragter für Datenschutz und Informationsfreiheit.

Rat für Sozial- und Wirtschaftsdaten RatSWD (2018). *Forschungsdatenmanagement in den Sozial-, Verhaltens- und Wirtschaftswissenschaften – Orientierungshilfen für die Beantragung und Begutachtung datengenerierender und datennutzender Forschungsprojekte* [Research data management in the social, behavioral and economic sciences - guidelines for the application and assessment of data-generating and data-using research projects] (2nd Edition). RatSWD Output 3(5) [Online Document]. https://doi.org/10.17620/02671.7

Rat für Sozial- und Wirtschaftsdaten RatSWD (2020). *Datenerhebung mit neuer Informationstechnologie. Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz [Data collection with new information technology. Recommendations on data quality and data management, research ethics and data protection]*. RatSWD Output 6(6) [Online Document]. https://doi.org/10.17620/02671.47.

Schönbrodt, F. D., Gollwitzer, M. & Abele-Brehm, A. (2017). Der Umgang mit Forschungsdaten im Fach Psychologie [Handling of research data in psychology]. *Psychologische Rundschau, 68*, 20-35. https://doi.org/10.1026/0033-3042/a000341

Sim, I., Stebbins, M., Bierer, B. E., Butte, A. J., Drazen, J., Dzau, V., Hernandez, A. F., Krumholz, H. M., Lo, B., Munos, B., Perakslis, E., Rockhold, F., Ross, J. S., Terry, S. F., Yamamoto, K. R., Zarin, D. A., & Li, R. (2020). Time for NIH to lead on data sharing. *Science, 367*(6484), 1308-1309. https://doi.org/10.1126/science.aba4456

Spindler, G. & Hillegeist, T. (2011). Rechtliche Probleme der elektronischen Langzeitarchivierung von Forschungsdaten [Legal problems with the electronic long-term archiving of research data]. In S. Büttner, H.-C. Hobohm & L. Müller (Eds.), *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock & Herchen.

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Members of the DGPs Commission "Open Science":

Mario Gollwitzer (Chairman), Ludwig-Maximilians-Universität Munich

Andrea Abele-Brehm, Friedrich-Alexander University Erlangen-Nuremberg

Christian Fiebach, Goethe University Frankfurt am Main

Anne Scheel, Eindhoven University of Technology

Felix Schönbrodt, Ludwig-Maximilians-Universität Munich

Ulf Steinberg, Technical University of Munich

Other contributors:

Maximilian Frank (official representative of the AG Open Science in the PsyFaKo)

Roland Ramthun (ZPID Leibniz Institute for Psychology Information Trier)

The original (German) version of this text is available here: https://doi.org/10.31234/osf.io/hcxtm. The English version, which has been drafted by Lisa Trierweiler (ZPID Trier) in cooperation with the DGPs Commission "Open Science" is available here: https://doi.org/10.31234/osf.io/24ncs.