4a Working Paper 2024

KonsortSWD
Consortium for the
Social, Behavioural, Educational
and Economic Sciences

Data access

Introduction to the topic of data access in the social, behavioural, educational, and economic sciences in research data centres

Ute Hoffstätter, Monika Linne

September 2024

www.konsortswd.de

Data access

Introduction to the topic of data access in the social, behavioural, educational, and economic sciences in research data centres

Ute Hoffstätter,[1] Monika Linne[2]

September 2024

DOI: 10.5281/zenodo.13768710

[1] German Centre for Higher Education Research and Science Studies (DZHW)
[2] RWI - Leibniz Institute for Economic Research – most of the time spent writing the paper: GESIS – Leibniz Institute for the Social Sciences

Note: This document has been translated from German (https://doi.org/10.5281/zenodo.7347064), therefore some of the references are only available in German. It focuses on research data infrastructure in Germany.

We thank for valuable advice (in alphabetical order): Daniel Buck, Dr. Andreas Daniel, Alexia Meyermann, Dr. Kati Mozygemba, Dr. Pascal Siegers, and Dr. Jonas Recker

Abstract

The central task of research data centres (RDCs) is to facilitate access to data for secondary use alongside the archiving of research data and related activities, including data documentation and data curation.[1] This article therefore supplies (prospective) RDCs and other research data infrastructures in the social, behavioural, educational, and economic sciences with essential information on the various options they have to offer data access paths to digital resources. To this end, the various data access paths are presented, which include downloading data, different variants of remote access, and on-site data access, and the properties that go with them as well as services for archiving and publishing. Aspects relevant to choosing a data access path are also highlighted in this context. One aspect includes the costs incurred by the various data access paths. The Five Safes model is utilised to explain the various parameters of data access and to illustrate the interrelation between these parameters.

The article also covers those characteristics of data that determine how open or restricted access to them can be or whether it is necessary to implement anonymisation measures. Data catalogues or research systems help with target group-specific access to data. They can provide information on access authorisations for data using access categories or by allocating standardised metadata.

Moreover, the article at hand points out the legal regulation that needs to be observed. They define which category of persons should be authorised to re-use research data, for which purpose the data may be used, and how this can be monitored. This is where terms of use, licenses, and data use agreements come into play, which, transparently and unambiguously, determine the research data's possible purposes of use and the terms under which they may be used.

A fundamental principle underlying sustainable access to research data are the FAIR principles, the application of which aims to facilitate the findability, accessibility, interoperability, and reusability of digital resources. For this reason, this paper will step-by-step present various tools and application examples of the practical implementation of the FAIR principles and thus relevant measures for sustainable data access, including options for obtaining persistent identifiers (PIDs), choosing appropriate licenses, or the relevance of schemes for standardised metadata. The content of this paper and the (FAIR) application examples concretely refer to basic measures of basic data access and thus represent an introduction to the topic.[2]

Keywords: Data access, data access paths, FAIR principles, RDCs, Five Safes

[1] Frequently, data are only made available in a restricted way, for example, only for certain user groups or purposes of use (e.g., scientific research).

[2] Since the article at hand is introductory in nature, further and more in-depth research data management practices in research data centres cannot be considered here.

Table of contents

1. Introduction

In the spirit of a scientific practice steeped in open science, research data should be described transparently and made available for further use as openly as possible. This facilitates the re-use of data and thus scientific progress and, ultimately, the progress of societies. Data access and its related processes play a significant role in the open-science practice. Within the various scientific disciplines, data access is facilitated to varying degrees and in different ways. This is due to the varying characteristics of those data, primarily the data's personal nature or the way copyright and other licensing issues are assessed. The article at hand is designed to supply basic guidance for research data centres (RDCs) to facilitate data access, specifically in the social, behavioural, educational, and economic sciences. The aim of this guide is to provide a general introduction to the topic as well as key information and best practice examples for designing possible data access paths in the context of setting up an RDC.

To this end, concrete options for implementing data access within RDCs are highlighted. These options for implementation are supplemented by the FAIR principles[3], which formulate key principles for research data management (RDM) that should be taken into account when setting up such services and research data infrastructures. By choosing this approach, the aim is to create a basic understanding of the topic and highlight concrete pathways for the implementation of RDM services based on the FAIR principles.

2. Data provision

2.1. Data access paths

There are different options for making research data available for re-use. Both the type of the data and the purpose of re-use are crucial for choosing a data access path. In the following, we will detail the most common data access paths and their special features. We will start with the least restrictive data access path. The procedures that we highlight after that are suitable for sensitive research data because technical and organisational measures facilitate monitoring data access and data sharing.[4]

3 https://www.go-fair.org/fair-principles/

4 For a current overview of data access paths at RDCs that are accredited by the German Data Forum (RatSWD), see its 2019 Activities Report (RatSWD, 2020b, p. 32ff).

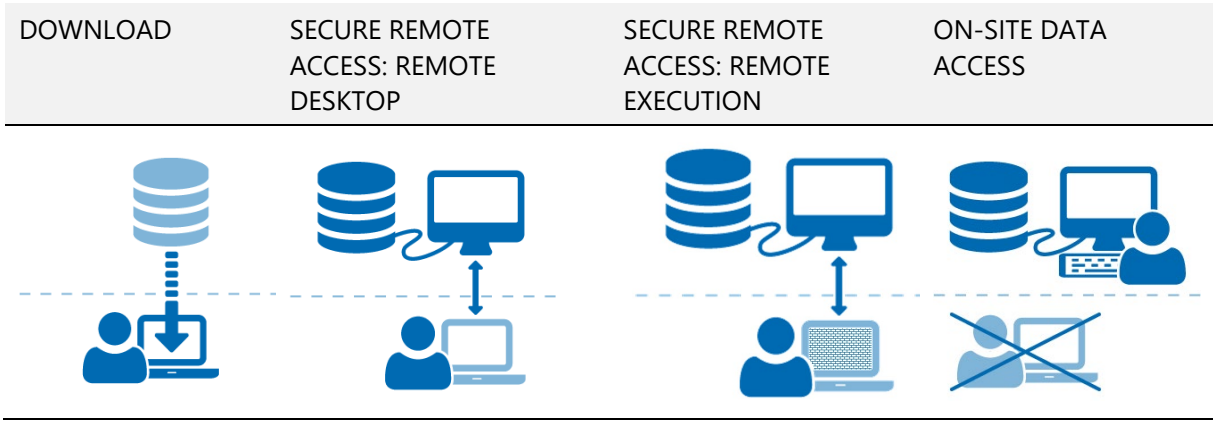

Figure 1: Data access paths
Source: https://www.fdz.dzhw.eu/en/data-usage, own diagram

Figure 1 shows the most frequently used data access paths, ranging from the least restrictive (download) to the most restrictive (on-site data access). They are presented in the following section and their consequences for data management and data users are described.

2.1.1. Download

The most open way of accessing data is *downloading* the desired data and any relevant documents from a data catalogue to the data user's terminal device. Depending on the data access strategy, this download can be completely free or require prior registration, authentication[5], or concluding a data use agreement. To ensure secure access, current security standards include encrypted transmission or two-factor authentication.[6]

2.1.2. Secure remote access

With more restrictive data access paths, the research data remain on RDC servers or servers of service providers but enable data users to access the data stored there. The umbrella terms *secure remote access*[7] or *remote access* refer to a variety of procedures in which off-site access is made possible, i.e., from the data user's workplace (D. H. Schiller et al., 2017, p. 7; D. Schiller & Welpton, 2015). In addition to data storage, this means that the data processing, too, takes place on RDC servers and that appropriate software for data analysis (e.g., Stata, SPSS, MAXQDA) must be made available there. This type of data access requires a particularly protected internal IT infrastructure that comprises a shielded network of virtual machines.

[5] Here, the required personal data should be kept to a minimum. Registration itself should be possible electronically and at no extra charge for the user.

[6] The SOEP is an example of this https://www.diw.de/en/diw_01.c.601584.en/data_access.html#c_741351, also the Research Data Centre of the Robert Koch Institute https://www.rki.de/SharedDocs/FAQ/FDZ/FAQ-Liste.html#FAQId13465420 (German only).

[7] For an extensive account of the various remote access methods, see D. H. Schiller et al. (2017) and RatSWD (2019). They also describe state-of-the-art approaches like "RDC-in-RDC".

Data provision using the *remote desktop* method[8] enables users to log on to the RDC server using software, view the data, and analyse them with software made available on the server. The users cannot download or import data of any kind. All files that users wish to import, or export are checked first by the RDC (input/output checking).[9] Following those checks, input and output are provided to data users by the RDC staff.

The more restrictive variant, the *remote execution* method, which is also known as *controlled remote data processing* or *remote execution* in German-speaking countries, enables users to log on to the servers of an RDC via a software. Unlike remote desktop, they cannot view the data. In this way, data users may analyse particularly sensitive data, which cannot be disclosed to them fully, however, for data protection reasons. Scripts or syntax for modifying, preparing, or analysing the data are sent to the RDC, without being able to view the data, and are then run by the RDC. The execution of the scripts and the transmission of the analysis results are possible at varying degrees of automation. All the files that users wish to import or export are checked first by the RDC (input/output checking). Output checking is mostly done supported by software but is typically also checked intellectually. The RDC staff subsequently makes in and outputs available to the data users.

2.1.3. On-site data access

With this form of data access (also known in German-speaking countries as guest researcher workstations, or GWAPs, safe centres, or secure data centres), data access takes place on-site at the RDC. Alongside remote execution, this data access path is the most restrictive variant of data access regarding technical and organisational safeguards. As with secure remote access, data storage and data processing take place on a RDC's central servers, but users work on specially equipped on-site workstations at the RDC instead of their own workplace. These are typically equipped with a computer without internet access, functioning USB ports, hard drives, etc. At most RDCs, additional regulation is in place for these facilities, including a ban on photographs, copying text, mobile phones, laptops, and the like. Users can view the data and analyse them using the software provided. However, users may not import or export data. All data that users wish to import or export must be checked beforehand (input/output checking). The next stage of expansion for on-site data access can be to create networks of those access stations across various institutions.[10] The minimum standards for connecting on-site data access stations regarding room security and criteria regarding the technical environment, which were developed in the KonsortSWD pilot project RDCnet, can also be applied as guidelines to conventional means of on-site data access (Murray & Goebel, 2022, p. 11ff).

Generally, researchers should choose the least restrictive data access path that is possible for the research data in view of their sensitivity in terms of data protection and research ethics

[8] In the English-speaking world, the term *(virtual) data enclave* tends to be used more.

[9] Output checks can take on different forms, see RatSWD (2019) for examples.

KonsortSWD is working on a pilot project aimed at connecting existing guest researcher workstations with each other (https://www.konsortswd.de/konsortswd/das-konsortium/services/rdcnet/). Access to data through the research data centres of IAB and GESIS is already connected via the IDAN project (https://www.gesis.org/en/services/processing-and-analyzing-data/research-visits).

issues. Additionally, it is necessary to consider resource issues—more restrictive data access paths often entail higher costs and require more effort for processing.[11] When planning multiple data access paths, it should be found out whether the same technical environments can be used for different data access paths. This would be conceivable, for example, for remote desktop and on-site data access, which saves costs. It should also be kept in mind that a once-built infrastructure requires secure future funding since the data are prepared for a certain access path that is in line with a data provision strategy and cannot simply be made available through other infrastructures.

Currently, only individual RDCs collect fees for data use (that do not cover costs). Typically, data use in Germany has been largely free of charge (see D. H. Schiller et al., 2017). However, regarding data use and data ingestion, there are discussions and some practical implementations of cost-sharing arrangements for data users and data providers. Cost-sharing is conceivable for cost-intensive data access paths (including remote execution, remote access), individual effortful data preparations for individual data users or for costly data ingest processes or long-term preservation. If RDCs provide a transparent cost model, reimbursement of these costs can be requested from the data providers or in turn from funding agencies, if applicable.[12] One way to be better able to calculate costs for a data provision infrastructure is to consult external service companies.

Alternatively, facilities without an appropriate research data infrastructure, where establishing a dedicated research data centre is not (yet) an option, can use existing services for archiving and publishing as (e.g., the generic repository RADAR[13], SowiDataNet for social and economic data[14], or one of the research data centres accredited by the German Data Forum (RatSWD)[15], etc.). The admission of data is subject to various conditions, which must be enquired about at the facility in question.

2.2. Choosing a data access path

The path through which the data are made available depends largely on the nature of the data themselves, for example, how sensitive they are (e.g., regarding their personal nature), which legal aspects the data entail (e.g., copyright and licensing rights, informed consent), aspects of research ethics (e.g., possible mental or economic harm to the respondents), and also the amount of data. If no legal or ethical reasons for curbing data access exist, they should be made available as *open data*[16].

[11] For more information on the costs of remote access environments, see D. H. Schiller et al. (2017, p. 22).

[12] The German Research Foundation, or DFG, for example, may take over "costs incurred specifically for a project in order to gain access to research data or to process and prepare the research data generated by the project in such a way that it can be used by others, as well as costs incurred by the transfer of data to a public repository. This includes personnel expenses for the preparation or transfer of data to existing repositories, as well as any software and hardware required for this purpose.", see information on the DFG resources available: https://www.dfg.de/en/research_funding/principles_dfg_funding/research_data/resources_available/index.html

[13] https://www.radar-service.eu

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

[15] https://www.ratswd.de/forschungsdaten/fdz

[16] https://www.forschungsdaten.org/index.php/Open_Access (German only)

Data from the social, behavioural, educational, and economic sciences are often subject to greater restrictions since personal data often require a higher level of technical and organisational safeguards for data access.

The protection of sensitive research data can be achieved through a variety of safety mechanisms. The portfolio approach of Desai et al. (2016) offers a helpful overview. This portfolio approach comprises the Five Safes dimensions (D. H. Schiller et al., 2017, p. 5f):

- safe people (people with appropriate training): Can the researchers be trusted to use them in an appropriate manner?
- safe projects (reviewed projects): Is this use of data appropriate?
- safe settings (technical environment): Does the access facility limit unauthorised use?
- safe outputs (controlled results): Are the statistical results non-disclosive?
- safe data (anonymised data): Is there a disclosure risk in the data itself?

As people with appropriate training (*safe people*), researchers have the knowledge, skills, and incentives to store and use data appropriately. Affiliation with an academic institution may be a good indicator, for example, as would be the acceptance of terms and conditions. Depending on the data, however, training on how to handle the data or the technical environment might also be reasonable. Reviewed projects (*safe projects*) take into account legal, moral and ethical considerations when using data (Desai et al., 2016, p. 8). In this case, for example, it is possible to enquire about the purpose of the project for which the data are required. Secure (technical) environments (*safe settings*) refer to the possibility of control over the data, ranging from a non-restricted download to on-site data access, to remote execution. Controlled results (*safe outputs*) refer to the results of data analyses. Particularly with less anonymised data, checking outputs is necessary. Anonymised data (*safe data*)[17] aim to minimise the risk of disclosure, which can vary depending on the degree of anonymisation.

The protection of data (and, with that, the persons under investigation) can be achieved by controlling persons, projects, settings, results, and/or the data. The Five Safes dimensions can help develop an adequate data access strategy. The approach is intentionally vague regarding operationalisation. However, it gives a good idea of how data can be deployed in a secure way by adjusting the individual dimensions (analogous to adjusting the sliders on a graphic equalizer).

[17] A multitude of methods exists for the process of data anonymisation itself (Müller et al. (1991); one of the most common is aggregation (e.g., aggregation of geographic information such as coarsening a place name to a country, or summarising age into age groups). In principle, the more aggregation, the more anonymisation, and the lower the risk of re-identification. At the same time, the information content of the data for research purposes also decreases. These concepts are described in Ebel (2015), Ebel and Meyermann (2015) as well as Eisentraut (2018).

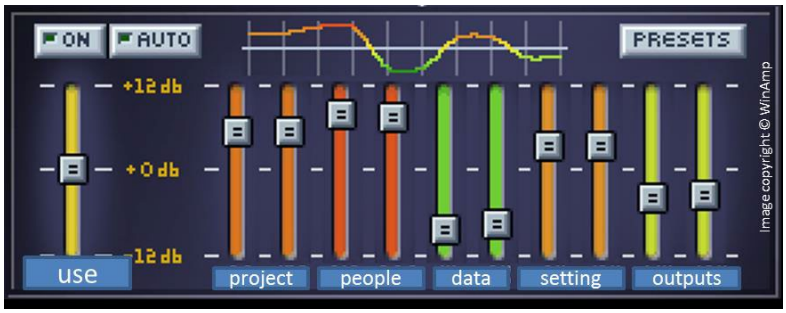

Figure 1: Five Safes (Desai et al., 2016, p. 5), based on McEachern, 2015

Figure 1 illustrates the concept behind the interplay of the dimensions. In this way, it would be possible to make available highly anonymised data (*safe data*) in an environment offering less safeguards of a more technical nature. By the same token, less anonymised data could be made available through more restrictive data access paths (*safe settings*). If users received training in advance (*safe people*), more sensitive data could be deployed.[18] Depending on the data type and data sensitivity, a different data strategy might be more appropriate. This strategy for the data or for entire series of studies should be developed in collaboration with the in-house data protection officer.

The following Table 1 shows examples of how these Safes can be designed.[19] The trustworthiness of a project (*safe project*) can be assessed by requesting information on the research endeavour during the registration or application process.

[18] Some RDCs require users to undergo training in advance, while others view academic affiliation and/or the acceptance of detailed terms and conditions as sufficient.

[19] The operationalisation of academic affiliation (*safe people*) for scientific use files is to be understood only as an example and not as a prerequisite (they can, for example, also include external PhD students that are employed outside of academia).

	SAFE DATA / DEGREE OF ANONYMISATION	SAFE SETTING / SECURE (TECHNICAL) ENVIRONMENTS	SAFE PEOPLE / PEOPLE WITH APPROPRIATE TRAINING	SAFE OUTPUTS / CONTROLLED RESULTS	SAFE PROJECTS / REVIEWED PROJECTS
1	Public Use File (PUF)	Download	No verification	No control, possibly terms of use	No control, possibly some additional information
2.1	Scientific Use File (SUF)	Secure download	Academic affiliation	No control, possibly terms of use	Requesting information
2.2	Scientific Use File (SUF)	Remote access through remote desktop	Academic affiliation	Input / output checks, possibly terms of use	Requesting information
3.1	Secure Use File (SecUF)[20]	Remote access through remote desktop in a safe room (connected GWAPs)	Academic affiliation, training	Input / output checks, possibly terms of use	Requesting information
3.2	Secure Use File (SecUF)	Remote access through remote execution	Academic affiliation, training	Input / output checks, possibly terms of use	Requesting information
3.3	Secure Use File (SecUF)	On-site data access	Academic affiliation, training	Input / output checks, possibly terms of use	Requesting information

Table 1: Overview of exemplary combinations of the Five Safes
Source: Own adaption based on D. H. Schiller et al. (2017, p. 5)

For more considerations and practice examples, related particularly to remote access solutions, see D. H. Schiller et al. (2017).

The data access strategy should be communicated in a transparent way, for example, by providing a document or information on the RDC website. Further, the terms of use[21] or a contract should be accessible so that users can look at the terms of use in advance.

2.3. Data

Data that are taken in by an RDC can typically not be made available for secondary use immediately. They must first be prepared, anonymised, and enriched with standardised

[20] For a definition, see section Dataa.
[21] These may differ depending on the data access path but also on individual datasets.

metadata[22]. The data are often made available to the users as standardised data bundles[23]. In some cases, however, the data are compiled on demand at the request of the user.[24]

The data, in turn, can be prepared in different ways—depending on the target group and the purpose. The following description is based on the terminology used to differentiate between EU data[25] (D. H. Schiller et al., 2017, p. 4).

- Public Use File (PUF)
- Campus Use File (CUF)
- Scientific Use File (SUF)
- Secure Use File (SecUF)

Public Use Files (PUF) are data that can be made available to all members of the public without any legal or ethical reservations (*open data*). In the field of social sciences, these often include highly anonymised data or structure files. *Campus Use File (CUF)* refers to data that are only made available for university teaching. These data, too, are often highly anonymised. *Scientific Use File (SUF)* as well as *Secure Use File (SecUF)* refer to data that are only made available for scientific research purposes. SecUF feature a low degree of anonymisation or are merely pseudonymised. In practice, the terminology is not equally widespread—SecUF and SUF are often used interchangeably because both are made available for scientific purposes.

RDCs and data providers decide together which type of data can or should be made available by the RDC. Legal and ethical aspects must be kept in mind here, including, for example, informed consent, i.e., the permission granted by the respondents to process the collected data, usually obtained before data collection, e.g., a survey. Provided there are no legal or ethical restrictions, it should be considered publishing the data in a Public Use File format. However, *informed consent* often only encompasses permission to share the data for scientific re-use[26]. In principle, several variants can be derived from one dataset to meet the needs of different target groups, a Campus Use File for university teaching, say, and a Scientific Use File for research purposes.

2.4. Data catalogues and data access

Once data have been created, they can be made available via an order system—often called data catalogue or research system. Such data catalogues contain descriptions of research data using metadata. Metadata feature central information on the research data, providing, on the one hand, potential data users with the necessary details on the research data and, on the other hand, facilitating the automated exchange of information between data catalogues. In the social, behavioural, economic, and health sciences, the Data Documentation Initiative (DDI)

[22] https://www.forschungsdaten.org/index.php/Metadaten (German only)
[23] A data bundle that was curated and prepared for general research purposes.
[24] The SOEP data, for example, can be compiled on demand: https://paneldata.org/
[25] COMMISSION REGULATION (EU) No 557/2013
[26] If personal data are to be passed on, the respondents must be explicitly informed and give their permission to this as part of obtaining their informed consent.

provides an internationally used, subject-specific metadata standard, which, due to its complexity, is well able to describe the entire data life cycle using XML, among others.[27]

Having been in operation for many years, many RDCs developed their own data catalogues to publish and publicise their research data. When describing data, it is essential to follow a metadata standard to ensure that the data and the metadata are exchangeable with other systems, thus complying with the FAIR principles, among others. Moreover, it is important that metadata can be retrieved using interfaces (e.g., OAI-PMH, Rest API) (see section FAIR data access). For this reason, it can be sensible to use existing software solutions[28] that already meet several requirements, including versioning and standards (e.g., metadata standards such as Dublin Core, DataCite, DDI, schema.org), and are continuously developed, in parts, by a larger open source community. These include Dataverse[29], DSpace[30], CKAN[31] or Dryad[32]. These applications provide structured information on research data according to established (metadata) standards. Most importantly, data and metadata are structured using controlled vocabularies, ontologies, and thesauri, which are either of a generic nature or relevant to specific academic disciplines. Using open interfaces, this information can then be exchanged in an automated and machine-readable way. Quasi automatically, several aspects of the FAIR principles have thus already been covered (see section FAIR data access).

To ensure the FAIR principle of accessibility (see section FAIR data access), it is recommended to also use a standard for describing the access categories in a data catalogue. Widespread categorisation schemes, such as the CESSDA standard[33], recommend four such categories:

- Open access
- Access for registered users (safeguarded)
- Restricted access
- Embargo

Open access refers to the possibility of direct data access, possibly after accepting the terms of use via opt-in consent. *Access for registered users (safeguarded)* means that data access is only possible for users who have registered with the RDC. The data in these two access categories do not contain direct identifiers. However, there is a risk of disclosure posed by linking indirect identifiers, so that special attention must be paid to data protection concerns[34] with this form of open data access. *Restricted access* means that data access is only possible after applying for it and after following an individual review procedure. This access category typically applies

[27] https://ddialliance.org/

[28] https://www.forschungsdaten.org/index.php/Repository_Software

[29] https://dataverse.harvard.edu/ The user interface is so far only available in English. However, additional languages are to be added as part of the EOSC.

[30] https://dspace-cris.4science.cloud/

[31] https://ckan.org/

[32] https://datadryad.org

[33] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Access-categories

[34] For further in-depth reading on the subject, the following literature is recommended: Kreutzer and Lahmann (2021), Depping (2021) as well as RatSWD (2020a)

to sensitive data that possibly contain personal information. Research data from the last category, the *Embargo* category, are subject to a limited embargo period. During the embargo period, only the description of the datasets (metadata) is published. The data themselves are made accessible at a later point in time. To do justice to point **A** of the FAIR principles, accessibility, descriptions of access categories should be made publicly accessible and possible procedures for checking access authorisations should be transparent for data users. The terms of use associated with the use of data (see section Legal regulation for data access) should also be accessible in advance. Regardless of which access categories are in place or which technical data access path leads to data access, metadata should always be made openly accessible, even if the data themselves are not (e.g., for data protection reasons), or are only made available for secondary use at a later date. There is currently no harmonised nomenclature for access categories in the German or the international community.

To aid the implementation of the FAIR principle of Findability, it is highly recommended to allocate a persistent identifier (PID)[35] that refers directly to the source of the research data. This makes the data uniquely identifiable and citable (see also section FAIR data access). In the social sciences, Digital Object Identifiers (DOIs) are the most frequently used. To obtain a DOI, certain mandatory information must be provided, i.e., at this point of the documentation process, the data are supplemented with information based on a structured scheme. With a metadata schema specialised on research data, DataCite is the central point of contact for DOI allocation. As the German DOI registration agency for social and economic data, da|ra[36] is registered as an intermediary with DataCite. Koch et al. (2017, p. 10) define the following mandatory fields for registering a dataset with da|ra, which are deemed essential for ensuring findability (see Table 2):

[35] https://www.forschungsdaten.org/index.php/Persistent_Identifier (German only)
[36] https://www.da-ra.de/

DA|RA PROPERTY	EXAMPLE: ALLBUS	EXAMPLE: NEPS
Resource type	Dataset	Dataset
Title	German General Social Survey (ALLBUS) 1986 – non-response study	NEPS Starting Cohort 6: Adults (SC6 12.1.0)
Creators	GESIS – Leibniz Institute for the Social Sciences	Artelt, Cordula (Leibniz Institute for Educational Trajectories, Germany); NEPS, National Educational Panel Study, Bamberg (Germany)
DOI	https://doi.org/10.4232/1.1669	https://doi.org/10.5157/NEPS:SC6:12.1.0
dataURL	https://search.gesis.org/research_data/ZA1669?doi=10.4232/1.1669	https://www.neps-data.de/default.aspx?tabid=5247
Publication date	2004	2021-12-09
Availability	Download	Download

Table 2: Mandatory fields for DOI registration at da|ra

When describing research data, these mandatory fields, and the mandatory fields of the metadata scheme of the DOI registration agency DataCite[37] should be used in any case because they adhere to an international standard and can thus be interpreted and processed way beyond the own data portal—resulting in **F**indability and **A**ccessibility in accordance with the FAIR principles. Moreover, filling in additional fields (such as keywords, description, geographic coverage) is highly recommended (Koch et al., 2017, p. 11). DOI registration can be done manually using an input form or via an API. With some services, including Dataverse, CKAN, or Dryad, automated allocation of DOIs is already integrated, or configurable. Automated registration via an API increases the integrity of the information because it is less prone to errors. In the data portal, the information for citing the data should be documented transparently. Additionally, a registration with da|ra enables various search portals[38] to crawl and list the items of research data. As mentioned above, it is advisable for the sake of simplicity to use existing software solutions that integrate these metadata standards.

2.5. Legal regulation for data access

In the social sciences, research data are typically made available for re-use only for the purpose of scientific research. For this reason, the first step in these cases is to review the scientific nature of the endeavour in question. This can be done either by relying on self-reporting, or by having RDC staff check the information that was supplied. The research's scientific nature should be assessed based on the group of users and the purpose of the data use. This is most often operationalised using a person's affiliation with an academic institution (e.g., university, non-university research facility) (group of users), in the context of which the research project is carried out.[39] Moreover, special attention should be given to the intended use since, in

[37] https://support.datacite.org/docs/datacite-metadata-schema-v44-mandatory-properties

[38] Examples include https://datasearch.gesis.org, https://www.base-search.net/, https://data.mendeley.com/ or https://sociohub-fid.de/

[39] In the case of student theses, it is possible to base this on the affiliation of the supervisor of the qualifying person with an academic institution. However, affiliation with an academic institution should not be interpreted as a mandatory legal requirement but solely as an indicator of scientific use.

principle, a private individual could also be working in a scientific way and, for example, be involved in creating scientific output. If no legal or ethical restrictions are in place and the data can be made available to the public, there is no need to consider the scientific nature of the purpose of use.

For the concrete provision of data, terms of use between data users and the RDC, or, if applicable, the data providers must be agreed upon. For anonymised data, which contain no reference to personal data (any longer), it is not necessary to apply the regulations of the GDPR. However, it is necessary to take aspects of research ethics into account (e.g., possible mental or legal harm of study participants). In principle, standard licenses[40] for open data, e.g., or licenses for scientific use[41] can be applied as terms of use. Here, the RDC is responsible for determining the type and design of the regulation that data access is subject to. Nevertheless, certain legal frameworks (e.g., GDPR for personal data, legal obligations towards data providers) must be considered. With data that potentially contain personal information, it is advisable to draw up a written data use agreement, which then creates a trustworthy legal framework for accessing the research data. Here, it can be useful to regulate requirements to technical and organisational safeguards (e.g., encryption of flash drives, reporting obligations) that are necessary for ensuring secure and responsible data access. A harmonised data use agreement, based on the harmonisation of 20 contracts of RDCs accredited by the German Data Forum (RatSWD), was developed within KonsortSWD and may serve as a template (Schallaböck et al., 2022).

The terms of use may also vary within any one RDC. Standard licenses might be used for certain data collections, or, again, a data use agreement might be drawn up for data collections containing potentially personal data. Whatever way the terms of use are ultimately designed, they are of particular relevance for complying with the **A**ccessibility criteria enshrined in the FAIR principles. Who can access the data under which conditions must be transparent and comprehensible at any given moment. In view of the best-possible **I**nteroperability of different datasets, it is recommended to use a standard that is mapped in a machine-readable metadata format. This enables computer systems to detect which research data can be used further under the same conditions and by which target group.

A widespread open licensing model for standards is made available by "Creative Commons"[42], a non-profit organisation. However, in Germany, applying only a creative commons license to social science research data is not advisable. Limiting the use of the data to scientific purposes,

[40] http://ufal.github.io/public-license-selector/ or https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data

[41] It could, for example, be reviewed whether the "License for scientific purposes ("Scientific Use License")" by PsychArchive could be used or taken as a template for one's own license: http://doi.org/10.23668/psycharchives.4988. The terms of use for data download of GESIS – Leibniz Institute for the Social Sciences can also be used as a template: https://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/_bgordnung_bestellen/2018-05-25_Benutzungsordnung_GESIS_DAS.pdf (German only).

[42] https://creativecommons.org/licenses/?lang=de

for example, which is a frequent prerequisite for re-use in the social sciences and adjacent disciplines, is not possible. In this case, additional options would include enquiring about the scientific purpose or falling back on of those mentioned above, such as using a data use agreement.

3. FAIR data access

The FAIR principles have become a firm fixture for assessing the re-usability of research data across the borders of countries and academic disciplines. They are the subject of much of the current discourse in research and have been included in the European Commission's funding guidelines[43] for Horizon 2020[44], meaning that they must be considered when applying for that funding scheme. It is the aim of the FAIR principles to ensure best-possible management and re-usability of research data for humans and machines and for new usage scenarios within the scope of what is legally and technically possible. This, however, does not mean that every dataset can or should be accessible without restrictions and at no charge, but merely that the requirements and access paths for accessing data collections are improved upon and are described in a transparent way.

The acronym FAIR refers to handling research data. They are to be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. The article at hand will take a closer look at the area of accessibility. The other three areas will be explained, too, to create a better understanding and to be better able to position accessibility within the general model of the FAIR principles. Lastly, they are also explained because the four principles partly build on and interlock with each other. The findability of data, for example, is a prerequisite for accessibility.

Findability

In order to be able to access research data, it must be possible to find them first. In this way, the principle of findability is essential for data access. Before research data or other resources can be found, however, the conditions for findability must first be created. (Machine-readable) metadata and persistent identifiers play a pivotal role here. Metadata are a basic requirement for the automated identification of datasets or services. Allocating persistent identifiers such as Digital Object Identifiers (DOIs) ensure the long-term findability and citability of resources. This is because the identifiers refer to the resource itself and not the storage location, making datasets findable even when the URL has changed (Brase, 2009, p. 57).

Several requirements must be met to facilitate and safeguard the findability of resources in the long term. Firstly, a resource must be equipped with a globally unique and persistent identifier (PID). Digital Object Identifiers (DOIs) have established themselves as a standard for referencing

[43] https://www.forschungsdaten.info/themen/informieren-und-planen/foerderrichtlinien/
[44] https://ec.europa.eu/programmes/horizon2020/en

data and digital objects[45] in academic publications. DOIs enrich a referenced object with metadata in accordance with the metadata scheme of the respective DOI registration agency[46] (e.g., Crossref[47] or DataCite[48]), which is in charge of allocating DOIs. Once a resource has been assigned a persistent identifier, neither the resource not the identifier may be subsequently changed. In principle, it is possible to assign a PID to an entire study including all related materials. Another option is to assign PIDs to each individual component of the study (dataset, syntax, questionnaires, codebooks, publications, etc.).

There are several ways to obtain PIDs or DOIs. Resources can, for example, be published independently through a repository (e.g., datorium[49] oder RADAR[50]) or an online storage service (e.g., Zenodo[51] or Figshare[52]), where they are assigned a PID as part of the publication process. However, caution is advised with online storage services because uploaded resources are neither reviewed nor curated. For RDCs, it is therefore advisable to go through DataCite, or to become a member of a consortium of a registration agency (e.g., da|ra registration agency for the social and economic sciences[53]) in order to obtain DOIs for research datasets or other materials. By registering with DataCite or a DOI registration agency, the metadata will be transferred into other systems automatically, making them findable in many subject-specific portals. Data that are indexed in this way can be searched by a web crawler and more easily found compared to those that are merely mentioned on a project's website.

A further condition for keeping with the FAIR principles is to describe resources using comprehensive metadata. Precise and extensive metadata are indispensable for ensuring the findability of digital resources. Based on the information provided in metadata, resources are made machine-readable and are thus made findeable by computers. The more detailed the metadata of a resource are, the easier it is to access said resource (GO FAIR 2022[54]). Since web content is accessed and made available in an increasingly automated way, it is important for the metadata to follow certain standards that structure information and to be understood by computer systems. On the website of a project, research data might be described using, say, a large text field. This can be understood by humans but not computer systems. When documentation adheres to a machine-readable standard, this information can be found and disseminated by machines as well. Much like when assigning PIDs, metadata can be created for entire projects or for all the individual components.

[45] Other persistent identifiers (PIDs) include Uniform Resource Name (URN), Handle System (hdl), Persistent Uniform Resource Locator (PURL), Open Researcher and Contributor ID (ORCID), and Gemeinsame Normdatendatei (GND), known as Integrated Authority File in English
[46] List of current DOI registration agencies: https://www.doi.org/registration_agencies.html
[47] https://de.wikipedia.org/wiki/Crossref
[48] https://datacite.org/
[49] https://data.gesis.org/sharing/#!Home
[50] https://www.radar-service.eu/radar/de/home
[51] https://zenodo.org
[52] https://figshare.com
[53] https://www.da-ra.de
[54] https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/

A distinction must be made between generic and subject-specific metadata standards. Generic metadata can be created independently from an academic discipline. The metadata generator DataCite[55], for example, is suitable for creating generic metadata. Discipline-specific metadata, however, are better able to describe certain resources because they are tailored to the specifics of a certain discipline. The Research Data Alliance (RDA) provides a list of various metadata standards[56], which can help find a suitable metadata standard for certain resources. Moreover, RDA has created a list of tools for generating standardised metadata.[57]

Accessibility

Once research data have been found, it must be known under which conditions the data can be accessed. In the FAIR concept, however, "accessibility" does not equal open or free-of-charge. Good reasons exist why research data might not be accessible or only under restricted conditions. This is particularly true for sensitive or personal data. Unlike the CARE principles[58], however, the FAIR principles do not explicitly refer to moral or ethical issues regarding the openness of research data. The decision as to whether and to what extent research data are published lies with the primary researchers, or the requirements of the respective institution funding the research. Accordingly, accessibility here does not refer to "whether or not" but the "how", i.e., the exact description of the conditions under which the research data can be accessed. This includes the conditions under which the data can be re-used (Mons et al., 2017, p. 49ff).

The conditions under which the research data or another resource can be accessed should be described clearly and transparently in the metadata (see also section Legal regulation for data access). Retrieval of the data should be possible without having to use special or proprietary tools. Instead, it should be possible using standardised communication protocols with the help of which the (meta)data can be retrieved via their identifiers. The protocols must be open, free, and universally implementable (e.g., HTTP, FTP, SMTP). In the case that (sensitive) research data are only accessible under certain conditions, the protocol should be able to support authentication and rights management. Should the research data not be generally accessible or no longer available, the metadata must stay permanently available (see section Data access paths).

Interoperability

Data are considered interoperable if they can be combined with other datasets by humans or machines. Therefore, data should be exchangeable and interpretable across computer systems. This means that they can capture automatically whether the content of the data is comparable to that of other data without requiring specialised or ad-hoc algorithms, translators, or mappings. Interoperability typically means that a computer system has at least some

[55] https://dhvlab.gwi.uni-muenchen.de/datacite-generator/
[56] https://rdamsc.bath.ac.uk/
[57] http://rd-alliance.github.io/metadata-directory/tools/
[58] https://www.gida-global.org/care

knowledge of the data exchange formats of the other system. For this to happen and to ensure automatic discoverability and interoperability of datasets, it is crucial to

- use metadata based on controlled vocabularies, ontologies, and thesauri that are established in the scientific community.
- use persistent identifiers that can be used to refer to other datasets. Here, it should be stated exactly whether a dataset builds on another dataset, whether additional datasets are required to complete the data, or whether additional information can be found in another dataset. It is particularly important to describe the type of relationship between the datasets in order to be able to make out the intellectual connection between the datasets.
- use a metadata scheme that too is based on the FAIR principles.[59]

Reusability

Ultimately, the goal of the FAIR principles is to facilitate the re-use of research data. For this purpose, both the data and the metadata must be described in great detail so that they can be used for replication purposes or any other further processing. For this, too, meaningful metadata are crucial, which describe the context in which the research data were collected in addition to the data themselves. This includes, for example, extensive documentation of the methods and devices that were used for creating the data. The more detailed and meaningful the metadata are, the easier it is to re-use the data they describe.

Moreover, (meta)data must be published under a clear license, which regulates the conditions for re-use, for humans and machines, in a transparent way. For metadata, it is recommended to use an open license such as the CC0 license[60]. The terms of use of the research data can be regulated in different ways (see section Legal regulation for data access). Citability is essential for being able to clearly mark re-use of research data. Proper citation of research data is made possible by assigning persistent identifiers, as detailed above, which facilitate long-term traceability (GO FAIR 2022[61]).[62]

Now, tools have been developed for assessing how FAIR data are. See Pittonet Gaiarin (2020) for an overview of these. The F-UJI tool[63] (currently still under revision) is especially suitable for an assessment of the FAIRness[64] of datasets.

[59] https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/
[60] https://creativecommons.org/publicdomain/zero/1.0/deed.de
[61] https://www.go-fair.org/fair-principles/
[62] For gathering more in-depth knowledge of the FAIR principles, it is recommended to refer to Betancort Cabrera et al. (2020), who do an excellent job of elaborating on the implementation of the more general FAIR principles in the social, behavioural, and economic sciences.
[63] https://www.f-uji.net/
[64] The operationalisation is based on Devaraju et al. (2020).

4. Summary/Outlook

This guide provides a basic overview of data access in the social sciences and adjacent academic disciplines. Since data in this community cannot very often be made available as open data, it is important to consider various aspects of appropriate data access. It should be thought through beforehand which infrastructures to use—and, specifically, which data access paths and data catalogues—since the dependencies and structures resulting from that choice have consequences for the medium and long term. Further care should be taken when calculating costs, possibly with the assistance of other RDCs and their experiences. It should also be examined whether it is possible to use existing, internal technical infrastructures, thus exploring further opportunities for saving costs.

The Five Safes provide us with reference points on which measures should be considered for secure data provision. They include considerations on the possible group of users and purpose of use for which the data are prepared and made available, including for, say, scientific analyses or a broader spectrum of uses. Digital services for data catalogues can make data access significantly easier. Transparent terms of use and standardised access categories aim to structure data access and should be made available to data users up front. This process can be supported by using existing open-source software applications for data catalogues, as they prepare and present information according to internationally standardised schemes. Establishing data catalogues is made significantly easier by using open-source software, compared to using entirely in-house developments. This is not least because RDM open-source tools often take into account most of the FAIR principles, thus reducing the conceptual expenditure here. Other options for complying with the FAIR principles in terms of data access were highlighted based on examples of application.

Further steps can be considered in order to internally evaluate and externally demonstrate the quality of the work processes and the infrastructure of an RDC. Data users and data providers looking for a suitable data archive, for example, are often referred to quality seals that give some indication of an archive's trustworthiness. For RDCs, obtaining accreditation of the RatSWD[65] can be a first step to highlight externally validated quality assurance vis-à-vis the scientific community. The accreditation criteria aim at equal treatment of all eligible data users in terms of data access and act as proof of the reliability of data access towards data users. The Core Trust Seal[66] is a further option for documenting an RDC's trustworthiness from the perspective of data archiving and data use. Obtaining this certificate involves in-depth documentation and examination of different areas of an RDC from a technical and organisational perspective. However, the expenditure involved should not be underestimated (see Pegelow et al., 2021). It does, however, offer an opportunity to critically reflect upon and optimise internal processes and to become more well-known at an international level.

[65] https://www.konsortswd.de/datenzentren/akkreditierung/
[66] https://www.coretrustseal.org

References[67]

Betancort Cabrera, N., Bongartz, E. C., Dörrenbächer, N., Goebel, J., Kaluza, H., & Siegers, P. (2020). *White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences.* https://doi.org/10.17620/02671.60

Brase, J. (2009). Der Digital Object Identifier (DOI). In H. Neuroth (Ed.), *Nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung* (2nd ed., pp. 57–65). Hülsbusch; Univ.-Verl. Göttingen.

Depping, R. (2021). *Rechtliche Aspekte des Forschungsdatenmanagements: Eine Einführung*. http://kups.ub.uni-koeln.de/id/eprint/45599

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five Safes: designing data access for research* (Economics Working Paper Series No. 1601). https://doi.org/10.13140/RG.2.1.3661.1604

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., Vries, J. de, L'Hours, H., Davidson, J., & White, A. (2020). *Fairsfair Data Object Assessment Metrics.* https://doi.org/10.5281/ZENODO.4081213

Ebel, T. (2015). *Empfehlungen zur Anonymisierung quantitativer Daten*. Mannheim. GESIS – Leibniz-Institut für Sozialwissenschaften.

Ebel, T., & Meyermann, A. (2015). *Hinweise zur Anonymisierung von quantitativen Daten* (forschungsdaten bildung informiert No. 3). Frankfurt am Main. Verbund Forschungsdaten Bildung. https://www.forschungsdaten-bildung.de/get_files.php?action=get_file&file=fdb-informiert_nr-7.pdf

Eisentraut, M. (2018). Data Anonymization. In S. Netscher & C. Eder (Eds.), *GESIS Papers. Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research* (pp. 34–36).

Koch, U., Akdeniz, E., Meichsner, J., Hausstein, B., & Harzenetter, K. (2017). *Da|ra Metadata Schema.* https://doi.org/10.4232/10.mdsdoc.4.0

Kreutzer, T., & Lahmann, H. (2021). *Rechtsfragen bei Open Science - Ein Leitfaden*. Hamburg. Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky. https://doi.org/10.15460/HUP.211

McEachern, S. (2015). *Implementation of the Trusted Access Model ASSA Policy Roundtable*. http://rsss.anu.edu.au/sites/default/files/SMcEachern_researcherperspective_ASSANov2015.pdf

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Müller, W., Blien, U., Knoche, P., & Wirth, H. (1991). *Die faktische Anonymität von Mikrodaten*. *Forum der Bundesstatistik: Vol. 19*. Metzler-Poeschel.

[67] All online sources were last checked on 20 October 2022

Murray, N., & Goebel, J. (2022). *Vertragliche Grundlagen zur Teilnahme am RDCnet*. Berlin. KonsortSWD. https://doi.org/10.5281/zenodo.6358334

Pegelow, L., Jansen, M., & Neuendorf, C. (2021). Erwerb des Zertifikats CoreTrustSeal (CTS) durch ein Forschungsdatenzentrum im Bildungsbereich – Motivation, Umsetzung und Lessons learned. Advance online publication. https://doi.org/10.17192/BFDM.2021.1.8310

Pittonet Gaiarin, S. (2020). *Fair Assessement and Certification in the EOSC region*. European Open Science Cloud. https://doi.org/10.5281/zenodo.4486280

RatSWD. (2019). *Remote Access zu Daten der amtlichen Statistik und der Sozialversicherungsträger* (RatSWD Output 5 No. 6). Berlin. Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://www.konsortswd.de/wp-content/uploads/RatSWD_Output5.6_RemoteAccess.pdf https://doi.org/10.17620/02671.42

RatSWD (2020a). Handreichung Datenschutz.: 2. vollständig überarbeitete Auflage. *RatSWD Output*, *8*(6). https://doi.org/10.17620/02671.50

RatSWD. (2020b). *Tätigkeitsbericht 2019 der vom RatSWD akkreditierten Forschungsdatenzentren (FDZ)*. Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://doi.org/10.17620/02671.56

Schallaböck, J., Hoffstätter, U., Buck, D., & Linne, M. (2022). *Mustervertrag Datennutzung KonsortSWD*. KonsortSWD. https://doi.org/10.5281/zenodo.5828114

Schiller, D., & Welpton, R. (2015). Distributing Access to Data, not Data. *IASSIST Quarterly*, *38*(3), 6. https://doi.org/10.29173/iq122

Schiller, D. H., Eberle, J., Fuß, D., Goebel, J., Heining, J., Mika, T., Müller, D., Röder, F., Stegmann, M., & Stephan, K. (2017). *Standards des sicheren Datenzugangs in den Sozial- und Wirtschaftswissenschaften*. https://www.konsortswd.de/wp-content/uploads/RatSWD_WP_261.pdf https://doi.org/10.17620/02671.15

Publishing details

Contact:
Ute Hoffstätter
German Centre for Higher Education and Science Research (DZHW)
Lange Laube 12
30159 Hannover
Tel.: +49 511 450670-404
hoffstaetter@dzhw.eu

Monika Linne
RWI - Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen
Tel.: +49 201 8149-267
monika.linne@rwi-essen.de

September 2024

KonsortSWD Working Paper:
KonsortSWD, as part of the National Research Data Infrastructure, is expanding offerings to support research with data in the social, behavioural, educational, and economic sciences. Our mission is to strengthen, expand and deepen the research data infrastructure for the study of society. It should be user-oriented and consider the needs of the research communities. An important cornerstone is the network of research data centres that has been built up for more than two decades by the German Data Forum (RatSWD).

This series contains articles on research data management that are produced in the context of KonsortSWD. Articles that have been externally and double-blind peer-reviewed are marked accordingly.

KonsortSWD is funded by the German Research Foundation (DFG) within the framework of the NFDI – project number: 442494171.

DOI: 10.5281/zenodo.13768710

Citation suggestion:
Hoffstätter, U. & Linne, M. (2024). Data access. Introduction to the topic of data access in the social, behavioural, educational, and economic sciences in research data centres. KonsortSWD Working Paper 4a/2024. Consortium for the Social, Behavioural, Educational and Economic Sciences (KonsortSWD). https://doi.org/10.5281/zenodo.13768710